

26 August 1982

Life as a wonderful one-hoss shay

In the next 20 years at normal fertility rates, the US population will age dramatically. Are science and medicine prepared for the challenge and will they respond in the best possible way?

It is a general human failing, or a failing of our culture, to turn a blind eye to the fact that one day we will be old and infirm. The search for eternal youth brings with it avoidance of the inevitability of old age. But a few researchers in the United States have been asking in recent years whether old age must entail senility, illness and loneliness or whether through research, people who are now in their 30s, 40s and 50s stand a chance of living out their lives in relative good health until the very end. Unfortunately, most biomedical researchers and doctors are merely human, they share the human preference to not think about old age — let alone about how to cure it.

Two who have held out this new goal, Lewis Thomas, chancellor of the Memorial Sloan-Kettering Cancer Center in New York, and James Fries of Stanford University Medical Center, invoke a poem by Oliver Wendell Holmes, called the "wonderful one-hoss shay". The one-hoss shay — American slang for a small carriage — worked without a hitch until one day it gave out all at once and died. Holmes wished that death would be so kind to us and spare us the gradual loss of our hubs, tyres, springs and thills.

Thomas argues that basic biomedical research at the cellular level — the breakthroughs in Alzheimer's disease (presenile dementia) and work on the fundamental processes of memory, the ageing of cells, the possible switches that are key to vascular abnormalities and even cancer — may be found through a steady assault at the fundamental level in coming years. In 1978 he made a prediction: "I believe that the major diseases of human beings have become approachable biological puzzles, ultimately solvable. It follows from this that it is now possible to begin thinking about a human society relatively free from disease . . . What can we die of, if not disease?"

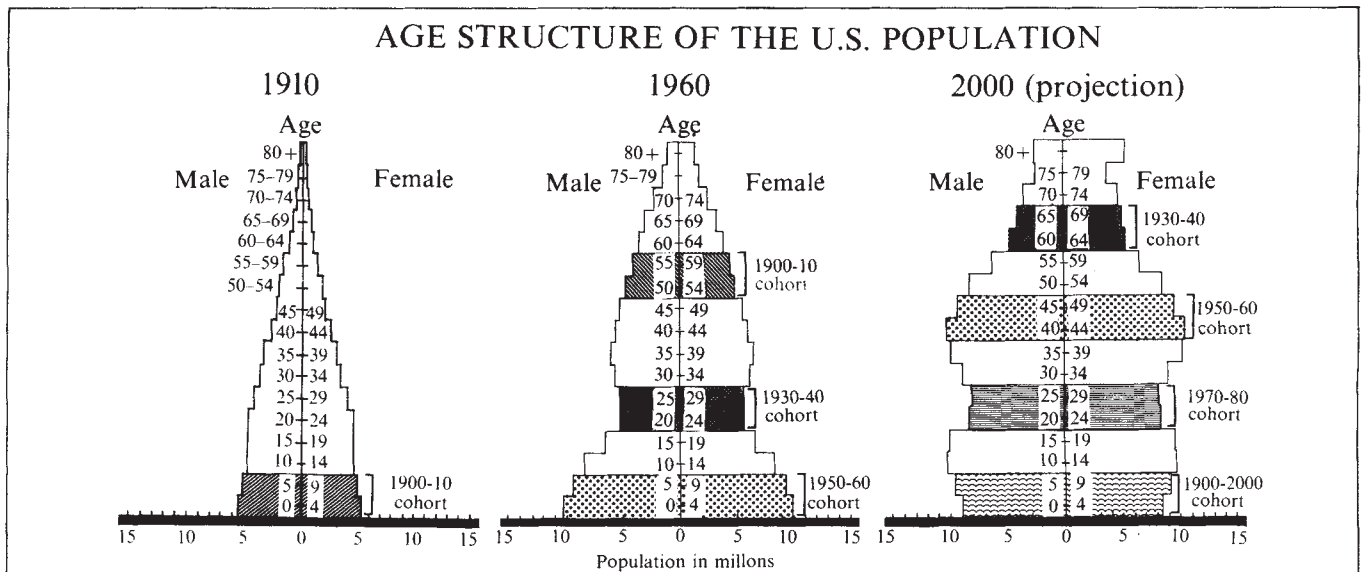
Fries has advanced the concept of the "compression of morbidity". Reduction in the acute illnesses from which people used to die continues to add to life expectancy in the United States. It is well known that a person celebrating their 65th birthday in 1953 could expect to live another 14 years; the person celebrating their 65th birthday in 1978 could expect to live another

16.3 years. On the other hand, Fries says that the life *span* of human beings will remain at 85. Medical science will enable more people to avoid premature death from an acute illness only to have them suffer from the familiar array of chronic diseases: cancer, emphysema and chronic bronchitis, vascular diseases, osteoarthritis and Alzheimer's disease.

The challenge of both arguments is the same: given the extraordinary increase in the numbers of old relative to the numbers of young in the future (see figure), can science make the lives of the old healthier, and relieve their dependence on younger family members, institutions, and the state? Although the challenge of the demographics is widely accepted, there are many arguments about how science and medicine should organize itself for this challenge.

One debate particularly among doctors is whether to establish geriatrics as a separate, licensed specialty, equivalent, for example, to pediatrics, cardiology and other medical fields. The debate going on in teaching hospitals and medical schools currently centres on whether a separate set of skills is required for dealing with the elderly and if so, what they are. In the United States, the problem is aggravated by the fact that a rather small proportion of doctors is family practitioners, and most are specialists. In the United Kingdom, which has a large proportion of old people, the problem is less serious, because most doctors are general practitioners, and treat the elderly as a matter of course.

Another debate is over whether or not to create new institutes within the US National Institutes of Health to focus attention on particular diseases of the elderly in order to attract money and political attention. There is already a National Institute of Aging, founded in 1976 and now consuming \$82 million a year. And some diseases of the elderly are already encompassed within other institutes; arthritis, for example, is covered by the National Institute on Arthritis, Metabolism and Digestive Diseases. But with the prospect of a new National Arthritis Institute being foisted upon him by Congress, Dr James B. Wyngaarde, director





of the National Institutes of Health, has been raising the kind of objection to focused institutes that is held by many serious biomedical researchers. Why fragment research, he argues, sensibly, just when the extraordinary interconnectedness of different diseases — for instance of the arthritis family with osteoporosis and systemic lupus erythematosus — is just becoming evident?

Typical short cycles for funding pose another obstacle to attacking the long-term, chronic diseases of the elderly. Most researchers (and doctors) prefer to work on problems that show quick, discernible results. But whether senile dementia is linked to nutrition or to early health, obviously cannot be answered in time for the next one- or three-year grant application.

Fortunately, more private benefactors are realizing that they can offer alternatives to traditional, government inspired approaches. Last week, Philippe Villiers, a Boston businessman, gave \$40 million to endow a new foundation dedicated to research and study of the elderly. Robert N. Butler, who is leaving the directorship of NIA after six years, will take a privately endowed chair as head of the country's first department of geriatrics, also privately endowed, at the Mount Sinai School of Medicine. But the profession would be better off, obviously, with some overall plan for how to address these new problems, rather than letting the whims of benefactors or the political process shape the future.

A start will be made soon by the National Academy of Sciences, which has just designated a chairman (Robert I. Berliner, dean of the Medical School at Yale) for its Committee on the Aging, now being formed under the auspices of the Institute of Medicine. It should be no surprise that the government did not see the need to fund an overall study (hence much of it will be paid for with academy funds) and that assembling the committee and subgroups has taken time. Ageing is not exactly the status-ridden scientific bandwagon that cancer was 10 years ago. But the scope of the study is very broad, ranging from the impact of the demographics on society to the needs of science and medicine, so the study will be a step in the right direction.

In his diary in exile, in 1935, Leon Trotsky wrote: "old age is the most unexpected of all things that can happen to a man". But there is nothing unexpected about the ageing of US population, and the future, greyer constituency for the results of research and for health care. If Thomas is right, biomedical scientists may have it within their power to make life more comfortable for these people — indeed for themselves in their old age. But to do this they must look the problems in the eye and face up to certain organisational and philosophical choices. One choice is for the US biomedical science community to follow the path of least resistance and set up new 'advocacy' specialities and institutes that would focus political attention on the problems of the elderly. The alternative, harder, but probably sounder, choice is to inject knowledge of, and concern for, the diseases of the elderly into all biomedical research to maximize the chances of the unexpected discovery that would lead, ultimately, to improved treatment or even cure of the diseases. Similarly, for the medical profession it would be better to make the needs of the elderly part of all medical training than to establish yet more specialities.

Language of love

French scientists are under pressure to communicate in French. But can they afford to?

The English edition of *Scientific American* used to sell only 5,000 copies a month in France. But the new French edition, *Pour la Science*, sells 50,000, its editor, M. Philippe Boulanger has claimed. So a mere change from one language to another increased communication tenfold.

This is what preys on the minds of those in France who wish to re-establish "French as a language of science". While French scientists rush to the microphone at conferences to address their American colleagues in English (see this issue p.784), and while more and more French scientists publish in English rather than

French (82 per cent of French papers were published in French in 1976, but only 67 per cent in 1980), a hunger — and a need — for knowledge among the rest of the French nation may go forgotten. Moreover, it is from this forgotten part of the nation that the next generation of scientists, technologists and technicians will emerge. What is more, it is imperative, from the point of view of a socialist government, that the same sector of the nation should, at the very least, be equipped to play an educated role in the democratic management of science and technology in the country. Meanwhile, many French primary scientific journals are declining — the best work being published in English — while few good textbooks are now published in French, and science is almost absent from radio and television. Is this the nation destined to become the technological leader of Europe?

One can therefore sympathize with M. Jean-Pierre Chevènement, Minister of State for Science and Industry, when he insists that French scientific culture must be based on the French language. His version of the French nation requires it. He does not decry international communications. There, he is the first to acknowledge, the universalism of English is a distinct advantage. Chevènement is determined to improve national communication and in pursuit of that goal Chevènement's scientist must look not only outwards, but inwards, towards his own countrymen and his own media.

But in practice, what can the minister do? He starts with the knowledge that most French scientists are against him. Professor Alfred Kastler, France's sole Nobel physicist, is a passionate Francophone, but he puts the Anglophone argument succinctly: "If we publish in French," the argument goes "the international scientific community doesn't read us. And if we speak in French at an international conference, most of the audience leaves the room". Briefly: to make contact with Americans, to keep in touch with his field, the French scientist must speak English. There is no simple solution to the dilemma that is posed for the French scientist sympathetic towards Chevènement's proposals but having to survive in a world dominated by the English language.

Professor Kastler has recommended "taking the offensive". But there is only one real offensive: to establish French science as a world force — just as the establishment of American science in the world (for whatever historical and economic reasons) led to the dominance of English. France can still dominate in one or two subjects — certain branches of mathematics, for example, and of social anthropology — but for the rest the battle will be arduous and the outcome must be gloomy.

Therefore, in most areas of science, French policy seems to grasp at straws. The provision of simultaneous translations at conferences has already been rejected as too expensive. The alternatives of partial translation by pre-print, transparencies or by a colleague (as now being suggested) seem artificial. In the matter of publication, to publish a paper in French as well as English is expensive and time-consuming; and will the editors of *Comptes Rendus* agree to become mere echoes of their counterparts on some English or American journal? Some have suggested more rigorous editing and selectivity in the French journals, and "vigorous marketing". This is a worthy proposal and can do no harm to the journals but if French scientists will not submit their best work to these journals anyhow, rigour and vigour are of little avail.

So is all lost for the French plan? Not entirely, insofar as early enthusiasms are now being tempered by reality. Certain small, but practical things are being done. Data banks and other forms of electronically stored information are being developed in French, for example, to make France independent of American data sources (there is a strategic imperative here also). And there are exercises to establish new, firmly edited journals in fields in which France does excel. Small things, perhaps, but the only practical ones. It is impossible to stage a complete revolution against English now. And as for French scientists: the measure of the success of the programme, in the end, will not be how many papers are published in French — but how many foreigners are forced to read them.

Vienna accord on remote sensing

Conference calls for bigger UN role

Vienna

New and expanded space-related activities, recommended to the United Nations by the Unispace-82 conference which ended last Saturday, should cost the United Nations \$3.2 million annually — almost four times the 1980–81 budget of the existing Outer Space Affairs Division of the UN Secretariat. These activities, the conference proposed, should be funded voluntarily by the member nations, and the UN General Assembly should rearrange its budget so that the increase in personnel costs may be absorbed within the available resources.

The proposed activities are largely information related, including consultancy services to assist developing countries in selecting space technology systems, study tours, seminars, and training fellowships, the creation of an international space information service, expansion of the existing space library and the establishment of a regular newsletter on space.

The need for such services became abundantly clear during the two weeks of Unispace-82; remote sensing has been constantly hailed as the answer to environmental and resource problems, and a bewilderingly large range of systems was on display for developing countries shopping for hardware or know-how. Yet the suppliers of these products were often first to admit that their data are simply not reaching those who most need the information. One representative of the UK Natural Environmental Research Council quoted the case of a remote sensing programme which identified icebergs heading directly for a Canadian off-shore oil-rig — with no organizational framework through which a warning could be transmitted to the rig at risk.

The conference report called for the formulation of international principles on satellite remote sensing which would allow the "sensed" state "timely and non-discriminatory access under reasonable conditions" to primary data relating to its territory.

The conference also expressed "real concern" over the approaching saturation of the geostationary orbit with communications satellites in certain frequency bands. Research into ways of facilitating the closer siting of satellites is "imperative", and the feasibility of using highly elliptical orbits for communications satellites should be re-examined. The International Telecommunications Union was

recommended to consider introducing into its future regulations a clause stating that a satellite owner was responsible for removing its hardware from orbit when it became inoperable.

Similarly, although the conference urged member countries to examine the feasibility of direct television broadcasting satellites for education in remote rural areas, it also urged them to consider the possibility of international or regional satellite space segments, shared by neighbouring countries.

In general, the conference confined itself to the specifics of space technology and its applications but there was also a full

discussion of the problem of preserving peace in space. After the three main committees of the conference had failed to reach consensus on this issue, Dr Willibald Pahr, the Austrian foreign secretary and president of the conference, convened a 15-member drafting group, called the "Friends of the president" which represented all regional groups and which finally, and with the conference in extra time, found an acceptable consensus wording calling on all nations, "in particular those with major space capabilities" to "contribute actively to preventing an arms race in outer space".

Vera Rich

Few winners in Australian budget

Sydney

The Australian government's latest budget, framed amid growing speculation about an early election, once again cuts back on funding for most areas of scientific research. The budget, released last week, was built around tax cuts and other handouts for the middle income earners. The government proved once again that they see very few votes in research funding.

Australia spends less than 1 per cent of gross domestic product on research, considerably less than most Western countries. The government provides about 70 per cent of this funding, which is an unusually large proportion by world standards. This year over half this money will go to the Commonwealth Scientific and Industrial Research Organization (CSIRO) and the rest to the other government agencies and the universities.

One of the few bright spots in the budget is the re-allocation of A\$49 million (£27 million) to a scheme that favours research projects with product-oriented goals. This

re-allocation came as a surprise because the scheme, funded under the Industrial Research and Incentives Act, was left up in the air last year when its budget was frozen after only A\$24 million was spent.

In the budget, basic research in universities has been given A\$19 million, around 4 per cent less in real terms than in 1981. This continues the steady decline which has gone on since 1975. The funds are channelled through the Australian Research Grants Scheme (ARGS) which now contributes only 15 per cent of the total national research spending compared with 24 per cent in 1975. If ARGS uses its reduced budget to maintain the top research groups, it will be unable to support about 100 projects that would otherwise have been accepted, according to Professor Max Brennan, chairman of the committee administering the scheme.

But there were some winners in the area of basic research. In astronomy, prior to the budget, two proposals were vying for patronage. One was the radioastronomers'

Canada's array poised

While Australian radioastronomers have just received the go ahead for their long baseline array of telescopes (see above) the fate of a Canadian counterpart is still in the balance.

The Canadian Long Baseline Array (CLBA) has already been approved in principle by the National Research Council (NRC) which has selected it over several competing big-science projects, such as the Canadian high energy electron ring and Starlab. A final proposal for the CLBA will probably be approved by NRC in September. The economic development committee of the Canadian cabinet will then have to decide whether it can afford the Canadian \$70 million needed to construct CLBA — with \$6.5 million per year running costs — in the midst of a recession.

The current proposal for CLBA calls for eight 32-metre dishes spread from British Columbia across to Newfoundland and one smaller dish in the North West Territories. Thomas H. Legg, the originator of the project, says the CLBA could be operating within four to five years of approval. The schedule is important, for the United States is planning its own very long baseline array that would stretch across the continent and include sites in Hawaii and Alaska. A proposal for building the system has been submitted to the US National Science Foundation (*Nature* 12 August, p.596) but whether it will be included in the US budget for fiscal 1984 is anybody's guess. There is a debate about the need for two separate systems of this size and both the Canadian and the US sides have thought about a possible link but nothing concrete has yet been proposed.

"Australia Telescope", the other the optical astronomers' "Starlab" project.

With the news that the Australian Telescope has been funded to the tune of A\$25 million, CSIRO can now proceed to develop a continental-scale, radio-linked interferometer network. When completed in 6 years' time, this should provide a Southern Hemisphere radio facility comparable to the most sophisticated interferometers in the north, complementing the activities in the optical and infrared bands of the Anglo-Australian Telescope.

The Australian Telescope will consist of a linear array of five 22-metre dishes at Culgoora in New South Wales, a 22-metre dish at Siding Spring (the site of the Anglo-Australian Telescope) and a 64-metre dish at Parkes. The total array will be equivalent to a dish 300 km across with a resolution of 0.1 seconds of arc — comparable to the US/European space telescope. It is also proposed that five other sites, covering much of the continent, can be radio-linked to the network, improving the resolution to one-thousandth of a second of arc.

The Starlab project has not been so fortunate. This joint Canadian, US, Australian scheme aims to place a 1-metre telescope in Earth-orbit by 1989. Australia's contribution was to have been the instrument package for the telescope. At this stage the government is not prepared to commit the full A\$28 million that would be necessary if Australia is to participate. But it is keeping the project alive by providing A\$3.3 million to local industry for some preliminary work.

Australian postgraduate research scholars were another notable group to gain in the budget. About one third of all full-time research students enrolled for higher degrees are supported on these scholarships. They have just been awarded a 50 per cent salary increase, presumably in recognition of the importance of their work as integral members of university research teams, and as Australia's future research scientists. Although this increase sounds impressive, the salary of a scholar has now only climbed from below the official poverty line to a generous A\$40 a week above (A\$6,850 a year). This is still less than half the average wage and no doubt a measure of the high esteem in which many Australian politicians hold Australian science.

Peter Hunt

US plant patent disputed

A fierce protest against the validity of a US patent dealing with plant breeding has been made public by Professor N.L. Innes, chairman of the British Association of Plant Breeders and a member of the staff of the British National Vegetable Research Station. The patent complained of was awarded in April this year to the Colorado based corporation Agrigenetics Research Associates, a seed firm with annual revenues of \$100 million.

The invention for which the patent has been awarded is described in the published version (US patent number 4,326,358) as a technique for accelerated production of new hybrid strains of plants and rapid commercial production of seeds from such hybrids. The patent claims that seeds of desirable new hybrids can be readied for marketing in as little as three years rather than the present eight to twelve years.

In conventional hybrid production, the plant breeder first has to breed two different homozygous plant lines from which a hybrid is produced and tested. Not only can it take many years to breed the homozygous lines but homozygous plants often produce low numbers of seeds.

In its essentials the invention covered by the patent starts with the crossing of any heterozygous plant — of which there is a great variety of good seed producers — with a heterozygous or homozygous partner. The hybrid offspring of such a cross will not be genetically identical but, on occasion, the plants will still be sufficiently similar to be worth testing as a potential crop.

If they have desirable crop qualities, the breeder then returns to the parent plants and propagates them, asexually, as clones. The large numbers of each parent so generated are then crossed to produce large numbers of hybrids, equivalent to those of the original cross of the individual parents.

The protest from the British Association of Plant Breeders (published in full on page 786) boils down to the assertion that the use and advantages of heterozygous parental plants as breeding stock are well known and that clonal propagation of individual plants is now a standard technique in plant breeding, so that the particular combination of the two principles for which a US

patent has been awarded must be obvious and thus not qualified for protection.

Even the combination of techniques described in the patent is very similar to that used in practice by, for example, British sugar beet breeders, says Dr Richard Macer, secretary of the British Association of Plant Breeders.

According to Rene Tegtmeier, of the US Patent Office, to which Professor Innes has sent a copy of the letter, a formal request for reexamination can be filed after a patent is issued, but only on the grounds of a prior patent or publication that was overlooked by the patent office in its original examination. Prior public use or sale is not sufficient grounds for reopening an already-issued patent. Even in the original examination of an application, Tegtmeier says, a foreign use would not bar patenting in the United States, although a foreign publication could.

"Any given detail or sequence may seem obvious, but the way they're put together may be original", so far as the patent law is concerned, says Dr David Padwa, chairman of Agrigenetics, who will shortly announce licensing terms that will be "fair and reasonable".

A second Agrigenetics patent, applying the techniques to a specific species, was recently allowed by the patent office and should soon be issued. Meanwhile Agrigenetics awaits the outcome of its application last January to the European Patent Office for a patent similar to the one issued in the United States.

Australian patents bill

Seeds of doubt

Canberra

The Australian government's first attempt to legislate for the protection of new plant varieties has blown up in its face. The Plant Variety Rights Bill, introduced a year ago and passed by the House of Representatives in April, is now the focus of a political storm. And the Senate has referred the bill to its Standing Commission on Natural Resources, a procedural device for postponing a decision.

The objective of the bill, of crucial importance in a country with a large agriculture industry, is to enable plant breeders to acquire the same kind of proprietary rights in new plant varieties as have long been available in some European countries. The present Patents Act, dependent as it is on the criterion of reproducibility, does not protect most plant varieties.

Five years ago, the Australian Agricultural Council (a political body representing federal and state ministers) recommended legislation on plant varieties protection with the objectives of stimulating the commercial plant breeding

List gończy



Komenda Wojewódzka Milicji Obywatelskiej we Wrocławiu poszukuje na podstawie listy gończego wydanego przez prokuratora rejonowego dla Dzielnicy Wrocław-Sródmieście Bolesława GLEICHGIEWICHA s. Adama i Marii z d. Majerane, ur. 30 IV 1919 r. w Warszawie, ostatnio zamieszkałego we Wrocławiu, ul. Norblina nr 21, zatrudnionego w Uniwersytecie Wrocławskim wydział matematyki, fizyki i chemii. Wymieniony podejrzany jest o to, że jako członek Komitetu Zakładowego NSZZ „Solidarność” w Uniwersytecie Wrocławskim nie odstąpił od kontynuowania działalności związkowej, mimo jej zawieszenia, lecz wziął udział w organizowaniu strajku w gmachu głównym tej uczelni oraz kierował jej przebiegiem. Jednocześnie ostrzega się, że za „ukrywanie” poszukiwanego lub dopomaganie mu w ucieczce grozi kara pozbawienia wolności od 3 do 15 lat.

"WANTED for continuing trade union activities under martial law, and for organizing a strike in Wrocław University" says this notice from a recent issue of the Wrocław daily *Gazeta Robotnicza*. Professor Bolesław Gleichgewicht, the subject of this notice, is a leading Polish mathematician, a former organizer of the clandestine "Flying University" and one of the founder-members of the Wrocław University chapter of Solidarity. He is now in hiding.

The notice includes a warning that the penalty for hiding or assisting the "fugitive" is from three to fifteen years loss of liberty.

industry, giving Australian farmers access to varieties developed overseas under the proposed bill's protection and enabling Australia to join L'Union Internationale pour la Protection des Obtentions Vegetales (UPOV).

As amended during the past year, the bill offers protection to Australian plant breeders producing plant varieties, including hybrids, that are novel, distinctive, stable and uniform. Existing varieties will not be entitled to protection.

Both government and opposition have been surprised at the controversy the bill has aroused. Its opponents include farmers, scientists, church groups, environmentalists, alternative life-stylers and consumer organizations. Many people employed by the federal and state governments on plant breeding fear that the bill, by making commercial plant-breeding more profitable, will give the governments an excuse to reduce support for plant breeding and also increase competition from the private sector, perhaps by the production of "cosmetic" varieties.

Even the claim that protecting plant varieties will stimulate the Australian private sector is disputed on the grounds that Australian farmers are at present only buying in one per cent of the seeds they sow each year — too little to generate much revenue. Some critics say that the most probable result will be to flood the market with seeds imported from overseas.

The fate of the bill is at this stage unclear. Hitherto, it had been thought that the fate of the bill would depend on the votes of the Australian Democrats, the minority party that holds the balance of power in the Senate and which sees its role as a watchdog over the machinations of the major parties. ("Keep the bastards honest" is its motto.) But several government senators now have cold feet about the bill.

One possibility is that nothing is decided until after the next election, particularly if that is called as early as the beginning of 1983. If there is an early election, everything will depend on which party is returned to Canberra. The present government might simply reintroduce the bill. The Labour Party, if elected, would probably let it die a natural death — and then find that it had to devise an alternative of its own.

Vimala Sarma

German nuclear power

Modest advance Heidelberg

Four new nuclear power plants at a go may seem like a boom but appearances are deceptive. The Federal German atomic power industry has problems. Although work began recently on the sites at Isar II (Bavaria) and Emsland, Lingen (Lower Saxony) and approval for Biblis C (Hesse) and Neckar-Westheim (Baden-Württemberg) seems little more than a for-

malism, all four reactors are of the same conventional high pressure light water, type — defensive planning that aims to expedite technical approval and confine local enquiries to siting and radioecology.

The electricity industry in West Germany is private, with a legal monopoly position based on laws dating from the 1930s. Most new power stations are financed by consortia, usually combinations of power companies and local government. The two-stage Federal-Land vetting procedure which keeps nuclear issues in political focus, stringent safety regulations, lengthy planning processes, and battles with environmentalist groups have turned the construction of atomic power plants in the West Germany into an obstacle race. Costs are now double those in France. Electricity prices are disappointingly high and industries now renegotiating 20-year contracts signed in the optimistic 1960s may consider importing from France. It is suggested that heavy industries may eventually emigrate to sites close to the French power plants. While West Germany has only 11 functioning nuclear power plants and 14 awaiting approval or under construction, *Electricité de France* has 24 functioning units and 26 in various stages of planning and construction.

Atomic power is controversial in West Germany: The CDU/CSU accuse the government of damaging the industry by ambivalence, imposing unnecessary controls, and dragging its feet on the reprocessing facility. The SPD is divided on the issue and the Greens (*Nature* 17 June) oppose use of atomic power categorically. Not only is the Bonn SPD/FDP coalition shaky and the SPD losing votes on the right to the CDU and on the left to the Greens, but the CDU itself is on the brink of a leadership conflict. With the Greens set to gain 10 per cent in the House election on 26 September and over 5 per cent in the Bavarian elections on 10 October, the major parties want to play down the nuclear power issue. For the time being there will be no decision on Biblis C which will add 1,300 megawatts to what at 2,500 megawatts is already the biggest atomic power complex in the country.

The Federal government participates financially only in prototype reactors. The fate of the fast breeder at Kalkar on the lower Rhine and the high temperature reactor at Schmehausen in the Ruhr await a meeting of the *Nuklearkabinett* on 31 October. In June this year, escalating costs led research and technology minister Andreas von Bulow to advocate halting both these projects. They were reprieved by Helmut Schmidt, chairman of the *Nuklearkabinett*, who insisted that for reasons of national prestige the projects should be retained.

Meanwhile the future of the Federal German nuclear industry looks as unclear as that of the Bonn government.

Sarah Tooze

Belgian nuclear fuel

Plant to restart

Waalre, The Netherlands

Eurochemic, a nuclear fuel reprocessing plant in Mol (Belgium) which was closed in 1974, is likely to begin work again. The international project shut down when Britain, France and Germany decided to go their own way, but on 2 July one of the two chambers of the Belgian Parliament voted to reopen the plant and it is expected that the Senate will do the same, at the earliest in October.

If the Senate agrees, a new fuel cycle company will be formed covering the whole nuclear cycle, with the Belgian government and the utilities taking equal shares. One or two subsidiary companies will take care of reprocessing and fuel fabrication. A new fuel will be produced: a mixture of uranium and plutonium.

The capacity of the reprocessing plant, which is now 60 tonnes per year, will probably be doubled. Without the plant, Belgian nuclear power stations would have had no place for spent fuel after 1985. A 120-tonne capacity at Mol also provides an opportunity to reprocess spent fuel from other countries.

The Belgian plant at Mol is now — after decontamination — cleaner than many experts thought possible, according to Dr Jacques van Gell. Radiation levels in the cells are only slightly higher than natural background levels, after 200 million curies have passed through them. "This is a world achievement", says Detilleux.

The reprocessing process will be changed at Mol, from the dissolving method to the mechanical chop and leach process. The existing fuel fabrication company Belgonucleaire, on the same site at Mol, will become part of the second subsidiary company and will produce plutonium for fast breeders but also for thermal reactors. Although Dr Detilleux considers that breeder reactors will not be needed for the next 15 years, using plutonium in conventional thermal reactors should give Belgium a more secure supply of fuel.

There has been considerable criticism of the Belgian vote in the Netherlands. The plant is only 15 km from the border, and after a number of ex-employees had told of incidents at the plant between 1966 and 1974, Dutch public interest groups protested against reopening and regional authorities asked for more information and for early warning systems in case of accidents. The Dutch under-minister for the environment, Mrs Ineke Lambers, was disappointed about the Belgian decision. Only the previous day, she had recommended in the EEC Council of Ministers that arrangements should be made for the European Parliament to settle such trans-border pollution issues. "This is a proof that such settlements are far away", she said.

Casper Schuurings

Language of science

French plans

The French government's campaign to re-establish French as a language of science is under severe stress — as thousands of foreign scientists descend on France for the 1982 conference season, attracted by the climate, the food and wine and, of course, the science. But these scientists mostly wish to speak English, not French, and there are not enough technical translators to go round (nor money to pay them.)

Breaking point came a couple of weeks ago at the 21st international conference on high energy physics in Paris, where even the French spoke English.

Government determination to push on with the language programme is unbowed. But there has been a shift of emphasis from the tone of the circular sent out to researchers last September, which spoke of the need for French scientists to make presentations "in our language". For at the high energy conference participants were faced with a questionnaire in four languages (French, German, English and Spanish) asking them what they thought of simultaneous translation, multilingual posters and other means of dealing with the language barrier. The point is that the government is now supporting not just French as a language of science, but also Italian, Russian, Japanese, Spanish, Arabic, Chinese and the rest. "We will do all we can to demolish the monopoly of English", said the programme director.

The objection to English is that — outside Anglophone countries — it leads to the creation of an artificial barrier between those scientists who can cope easily with the language and those who cannot, so that English is used as a matter of status and hence of isolation of one group from another. English thus militates against "democracy" in French science. Moreover, in France the government has taken an overriding interest in all means of increasing communication between scientists and industry, and this "snobbish" division is seen as a real barrier to industrial development.

Government determination, however, is not enough. Money is also required. To provide simultaneous translation for a number of languages at a major conference is expensive — about FF350,000 (£30,000) for a conference of 1,500–3,000 participants for three or four days, estimates MIDIST, the Mission Interministerielle de l'Information Scientifique et Technique in which the "French as a language of science" programme is based. Short of the necessary wherewithal, and of technically-trained translators who can deal with immunology on the one hand or high-energy physics on the other, MIDIST this year is resorting to experiments on means cheaper than simultaneous translation.

For a total cost of FF600,000 at 20 conferences over the next couple of

months, MIDIST will therefore try out:

- Simultaneous translation at plenary sessions, but not at working groups, which would be divided by subject and language.
- Distribution of a typed resumé of presentations in various languages.
- Bilingual transparencies, to be prepared before each talk.
- Paragraph-by-paragraph translation of a talk by a bilingual colleague.
- The résumé on request, by a participant, of the essential points of a talk in the language requested.

The results of these experiments will be tested in part by questionnaire, and may result in more positive action during the conference season of 1983. **Robert Walgate**

Astronomer on fast

Patras, Greece

Professor Leonid Ozernoi of the P.N. Lebedev Physical Institute in Moscow began a hunger strike on 17 August in the face of the refusal of the Soviet authorities to grant him and his family visas for emigration to the United States. His case was being widely publicized at the General Assembly of the International Astronomical Union at Patras, Greece which finishes on 26 August.

Professor Ozernoi is well-known for his work in high-energy astrophysics including theoretical studies of processes in quasars and the nuclei of active galaxies. He applied for emigration in June 1979. Following the application, his wife lost her job as a sociologist and has since been unable to obtain another position; he was barred from attending international conferences, removed from the board of *Letters to the Astronomical Journal of the USSR* and prevented from teaching. A paper of his was omitted from a book for which it was scheduled. Nearly 2½ years after applying to leave the country he was refused, his departure being "considered inexpedient at this time".

The Harvard-Smithsonian Centre for Astrophysics has invited Professor Ozernoi to work there, while the executive committee of the International Astronomical Union (IAU) will be considering whether or not to become involved officially at its next meeting in September.

A petition on behalf of Professor Ozernoi was circulated at the IAU Assembly. Addressed to E.R. Mustel, Chairman of the Astronomical Council of the USSR Academy of Sciences, it also concerned Dr Alpert of the Institute of Terrestrial Magnetism Radio Research and the Ionosphere of the USSR Academy of Sciences, who has experienced demotion and restrictions on his work since his application for exit in 1975.

Philip Campbell

Drug promotion

FDA steps in

Washington

"A new choice for freedom from arthritis pain . . . Coming soon from Geigy."

This advertisement, accompanied by a colour illustration of a dove flying from a pair of outstretched hands, appeared in medical journals last autumn. At first glance, it seemed much like the many other eye-catching advertising pieces that the drug companies invest in so heavily to promote their wares in their highly competitive business. The strange thing about this advertisement, though, is that nowhere is the name of the drug mentioned. In fact, the drug had not yet received approval from the Food and Drug Administration (FDA).

Over the past year or so, this tactic and others for the pre-approval promotion of drugs have become increasingly common. Recently, FDA has been trying to do something about it.

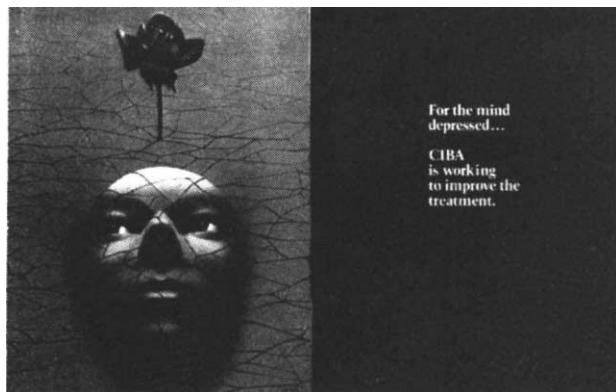
FDA has broad powers under the federal drug laws to regulate the labelling of prescription drugs it has approved. By extension, this has been applied to advertising and other promotional literature. Any advertisement for a prescription drug must, for instance, include complete information on side effects, precautions and proper dosage. FDA also scrutinizes headlines and claims made in promotional material for accuracy and proper balance. A false or misleading promotion can render a drug "misbranded" and liable for seizure — a remedy FDA has occasionally used.

FDA regulations also prohibit the promotion of a drug's safety or usefulness before approval, since it is the safety and usefulness that FDA is attempting to determine in its review of a new drug application. But what if the drug's name is not mentioned in pre-approval advertising? That was apparently what the drug companies asked themselves; and thus a loophole was opened.

Possibly the most effective of these pre-approval advertisements are of the variety shown opposite. A graphic motif is introduced in the pre-approval advertisement, which is carried over to the post-approval — and, the advertiser hopes, the favourable first impression is carried over as well.

No warnings, precautions or contraindications encumber the pre-approval advertisements. But neither does the drug's name, and that puts FDA in an awkward position. "Without the mention of a drug product, it becomes difficult for us to say that is an advertisement" under FDA regulations, says Kenneth Feather, acting chief of FDA's drug advertising regulation branch.

Now the agency hopes to set limits on pre-approval advertisements. Last month, FDA issued a list of "recommendations" aimed to stop what it sees as infringements of the spirit of the law.



An impressionistic pre-launch . . .

In particular, the agency said, it will object to any advertisements that "make claims of safety and efficacy even if the new drug is not specified". On the other hand, if the name is specified, FDA says, advertising copy may not suggest the drug's indications, uses, safety or effectiveness, nor may any other advertisements or even the graphics that appear. In other words, the advertisement can say only "Fribulate, coming soon from Wonder Drugs, Inc."

Pre-approval advertisements such as that shown here for Ciba's Ludiomil fall in a middle ground. The FDA says they are in general acceptable, but the dividing line is fine.

A case that came nearer to that line was Marion Laboratories' pre-approval campaign for Carafate, an antiulcer drug. The advertisements proclaimed the new, unnamed drug "a new era in ulcer therapy", a claim that FDA might well take exception to if it appeared in post-approval advertisements naming the drug.

Another loophole that concerns FDA is pre-approval promotion of drugs at scientific meetings. The FDA regulations specifically exempt seminars and scientific publications from its ban on representing a drug as "safe or useful" prior to approval; this is to allow a full exchange of scientific information concerning the drug. But according to Kenneth Feather of FDA,

some firms have been stretching their interpretation of this exemption. He says he has seen several "scientific exhibits" at meetings that had all the hallmarks of promotional exhibits of the "flashing lights variety".

At the recent Pan-American Congress of Rheumatology, for example, Lederle Laboratories, had an exhibit which offered to visitors who stopped by free computer photographs along with a flyer with the headline "Fenbufen: a highly effective non-steroidal anti-inflammatory drug with an unusually low incidence of serious GI complications". According to a letter that Feather sent Lederle, the exhibit was promoting Fenbufen, which the FDA has not yet approved for sale in the United States, although it is sold elsewhere.

After FDA objected, Lederle issued a statement that it "regarded the congress as an international meeting" and that "since requirements and practices differ from country to country . . . it does not appear equitable to adhere to those of any single country". FDA officials say, however, that it is nothing new for US laws to apply within the United States to US companies.

A similar incident occurred last year at the American Rheumatism Association meeting in Boston, where Pfizer Pharmaceuticals showed a film promoting Piroxicam, an anti-arthritis drug not yet

approved. FDA said that the film hardly qualified as a "scientific exchange" since it was shown at a cocktail and buffet reception sponsored by Pfizer and that there was virtually no scientific discussion. According to FDA, the only question from the audience following the screening was, "When will the drug be available?"

FDA's recourse in these cases is unclear. They may order the companies to stop an advertisement or not to show a film again, but this, says Feather, is "shutting the barn door after the horse has bolted", since most of these pre-approval promotions are designed to be one-shot campaigns. And the remedies available after a drug is on the market, such as seizure, do not apply beforehand.

The recent Oraflex case, though, may cast a chill over the industry's own eagerness to pursue aggressive promotional campaigns. In late July, shortly before sales of Oraflex were suspended following reports of 61 deaths among users in Britain, FDA notified Lilly that a press kit prepared for Oraflex contained false or misleading statements. FDA said it seriously understated the drug's side effects and implied that Oraflex had a unique ability to fight the underlying disease in arthritis patients. In fact, FDA says, Oraflex is just another non-steroidal anti-inflammatory agent, and any claims about its disease-fighting ability are based solely on speculative inferences from animal studies.

After such a hard sell, the backlash that came when the reports of adverse effects started accumulating was all the greater. If the drug is ever cleared for sale again, FDA will probably require stiffer labelling designed to counter the false impression that Lilly's press campaign created.

FDA officials have raised the spectre of stiffer labelling requirements to counter pre-approval claims as well, a threat that will probably weigh heavily in the industry's decisions on future promotions.

Stephen Budiansky

. . . succeeded by the complete story — and a name

A disease that affects millions of people throughout the world. The sufferers of which can truly be called . . .

The wretched of the earth: The depressed.

Ludiomil maprotiline HCl

A new generation antidepressant. Rises above the tricyclics. To provide safety with optimal effectiveness.

- fairly common side effects, such as dry mouth, constipation, and drowsiness;
- fewer severe anticholinergic side effects;
- symptomatic relief within the first week in some patients.

CIBA

Ludiomil maprotiline HCl

25 mg 50 mg

Indications: Ludiomil is indicated for the treatment of major depressive illness. It is also indicated for the treatment of anxiety disorders, such as generalized anxiety disorder, panic disorder, and agoraphobia. Ludiomil is also indicated for the treatment of chronic pain, such as neuropathic pain, and for the treatment of sleep disorders, such as insomnia.

Contraindications: Ludiomil is contraindicated in patients with a history of hypersensitivity to maprotiline or any of the components of the formulation. It is also contraindicated in patients with a history of severe liver or kidney disease, or in patients with a history of severe heart disease, such as congestive heart failure or a recent myocardial infarction.

Warnings: Ludiomil may cause drowsiness, dry mouth, constipation, and dizziness. Patients should be warned of these side effects and advised to avoid alcohol and other sedative drugs while taking Ludiomil. Ludiomil may also cause weight gain, which should be monitored in obese patients. Ludiomil may also cause changes in blood pressure, which should be monitored in patients with hypertension.

Precautions: Ludiomil should be used with caution in patients with a history of glaucoma, as it may increase intraocular pressure. It should also be used with caution in patients with a history of seizures, as it may lower the seizure threshold. Ludiomil should be used with caution in patients with a history of bleeding disorders, as it may increase the risk of bleeding.

Interactions: Ludiomil may interact with other drugs, such as alcohol, sedatives, and anticoagulants. Patients should be warned of these interactions and advised to avoid these drugs while taking Ludiomil.

Adverse Reactions: The most common adverse reactions to Ludiomil are drowsiness, dry mouth, constipation, and dizziness. Other adverse reactions include weight gain, changes in blood pressure, and changes in heart rate.

Pharmacokinetics: Ludiomil is absorbed rapidly and reaches its peak plasma concentration within 2-4 hours. It is metabolized in the liver and excreted in the urine. The half-life of Ludiomil is approximately 12 hours.

Pharmacology: Ludiomil is a tricyclic antidepressant that acts by inhibiting the reuptake of norepinephrine and serotonin into the presynaptic terminal. This results in an increase in the levels of these neurotransmitters in the synaptic cleft, which leads to an antidepressant effect.

References: 1. Ludiomil (maprotiline HCl) package insert, Ciba Ltd, Basel, Switzerland. 2. Ludiomil (maprotiline HCl) package insert, Ciba Ltd, Basel, Switzerland.

CIBA

CORRESPONDENCE

Patents and plant breeding

SIR — In the United States a patent (No. 4,326,358 of 27 April 1982) has recently been granted to Agrigenetics Research Associates Ltd of Denver, Colorado for the use of tissue culture (micropropagation *in vitro*) for the multiplication of parental plants to produce hybrid seed of crop varieties. The patented procedure is intended for the rapid development, evaluation and commercial production of hybrid seed from pairs of parental plants chosen for their specific combining ability and uniformity of the hybrid between them irrespective of their degree of homozygosity. The specific example describes the application of the technique to tomatoes. It is understood that equivalent patents in other countries have also been sought.

Both public and private plant breeders must feel concern that this patent purports to give rights over the use of techniques that have been part of the stock-in-trade of plant breeders for some considerable time and have already been used commercially. We would therefore like to draw attention to the conditions governing the validity of patents on plant tissue culture in the United States, as well as to evidence in the scientific literature on the widespread use of tissue culture techniques by plant breeders before the US patent application was made.

Bagwill¹ has summarized the validity of patents on plant tissue culture techniques in the United States. Among the key points made by Bagwill are:

- A person shall be entitled to a patent unless — (a) The invention was known or used by others in this country, or patented or described in a printed publication in this or a foreign country, before the invention thereof by the applicant for patent.
- (b) The invention was patented or described in a printed publication in this or a foreign country or in public use or on sale in this country, more than one year prior to the date of the application for patent in the United States.
- A patent may not be obtained . . . if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains.

In the light of Bagwill's remarks it is pertinent to observe that: (1) The essential principles and techniques in the patent were known and practised by others well before the patent application was filed. (2) The practical value of the particular combination of pre-existing principles and techniques outlined in the application had been recognized previously by others. (3) The value of the approaches outlined in the application to variety development and production were already obvious to any person having ordinary skills in the relevant disciplines.

The case for patentability is based on a number of premises which merit comment: *Utilization of hybrid vigour in developing crop varieties*: The value of hybrid vigour in

developing superior crop varieties has been recognized and exploited by plant breeders for more than 100 years. Beal was the first to recommend the use of hybrid varieties of maize; recognition of their commercial worth in other crops followed later. This principle is now universally acknowledged by biologists and is discussed in most textbooks on plant breeding².

Phenotype uniformity obtained from specific combination of heterozygous parents: The concept that first generation hybrids of acceptable phenotype can be obtained by crossing two heterozygous stocks has been known and used for a very long time. For example, varietal hybrids of maize were exploited commercially at least 150 years ago³. Controlled experiments on varietal hybrids of maize were reported in the 1870s⁴ and interest in them was revived in the 1960s⁴. The difference between a varietal hybrid and the type of hybrid described in the patent is one of degree rather than of kind.

In the description of the patented procedure it is noted that some, but not all, hybrids of pair-crossed heterozygous plants are sufficiently uniform in phenotype to be acceptable as commercial varieties. This is not a new idea. Differences in uniformity for important traits were noted among double-cross hybrids of maize more than 40 years ago⁵. To any person with a knowledge of genetics it is obvious that if there are differences in phenotypic uniformity among double-cross hybrids, they would also occur among hybrids produced by crossing heterozygous single plants because the same genetical principles apply in both cases.

Progeny testing and specific combining ability: In the patented procedure, selection among potential parents of hybrids is based on the field performance of their hybrid progeny. It acknowledges that the best way to exploit specific combining ability is to make and test the hybrid that will eventually be released as a new variety. The concept of exploiting specific combining ability in this manner is not new. It was the basis of Hull's⁶ recommendation that an inbred tester be used in developing new single-cross hybrids and is also an important feature of the full-sib selection systems developed by Hallauer⁷, Lonnquist and Williams⁸ and Hallauer and Eberhart⁹ for maize. The only significant difference between these methods and the patented method is the manner in which selected heterozygous plants are propagated.

Seed yields from heterozygous parents: The principle that greater yields of good quality hybrid seed are obtained if the parents of a hybrid are themselves heterozygous has long been known to breeders of cross-pollinated crops. It was embodied in Jones'¹⁰ proposal of 1918 that double-cross hybrids of maize should be used as commercial varieties. The same principle was later exploited in three-way cross hybrids and again in modified single crosses and sister line crosses². It has also been used in the production of kale hybrids¹¹ and is a concept used in the breeding of hybrid Brussels sprouts¹² and onions¹³.

Vegetative propagation: The vegetative propagation of heterozygous plants to provide adequate stocks of genetically identical plants for hybrid seed production is crucial to the procedure patented. However, the idea that cloning can be used to perpetuate and increase unique genotypes that cannot be maintained without genetic change by sexual methods of reproduction is common knowledge^{2,14-17}. Its particular use in the production of hybrid seed has also been repeatedly recommended^{2,11,14,18-25}. The special value of micropropagation *in vitro* as a potentially more efficient substitute for vegetative propagation *in vivo* is also generally acknowledged^{14,15}.

On the basis of our interpretation of Bagwill's paper¹ and the evidence presented here, we find it difficult as scientists to see how the conditions recited by Bagwill can have been satisfied in the process of obtaining this patent. We strongly urge the United States Patent and Trademark Office to re-examine this particular patent and if similar applications are made elsewhere to consult specialist opinion more widely before granting other plant patents that may similarly restrict the use of techniques and combinations of techniques that are common currency among plant breeders worldwide.

N. L. INNES

British Association of Plant Breeders,
The Plant Royalty Bureau Ltd,
Woolpack Chambers,
Market Street,
Ely, Cambs. CB7 4ND, UK

1. Bagwill, R.E. *Envir. exp. Bot.* **21**, 383-387 (1981).
2. Allard, R.W. *Principles of Plant Breeding* (Wiley, New York, 1960).
3. Sprague, G.F. & Eberhart, S.A. in *Corn and Corn Improvement* (ed. Sprague, G.F.) 305-362 (American Society of Agronomy, Madison, 1977).
4. Eberhart, S.A., Harrison, M.N. & Ogada, F. *Der Zuechter* **37**, 169-173 (1967).
5. Jugenheimer, R.W. *Agricultural Development Paper* No. 62 (FAO, Rome, 1958).
6. Hull, F.H. *J. Am. Soc. Agron.* **37**, 134-145 (1945).
7. Hallauer, A.R. *Crop Sci.* **7**, 192-195 (1967).
8. Lonnquist, J.H. & Williams, N.E. *Crop Sci.* **7**, 369-370 (1967).
9. Hallauer, A.R. & Eberhart, S.A. *Crop Sci.* **10**, 315-316 (1970).
10. Jones, D.F. *Connecticut Agricultural Experimental Station Bulletin* **207**, 1-100 (1918).
11. Thompson, K.F. *Agriculture, Lond.* **65**, 487-491 (1959).
12. Johnson, A.G. *Commercial Grower* **4014**, 831-832 (1972).
13. Erickson, H.T. & Gabelman, W.H. *Proc. Am. Soc. hort. Sci.* **64**, 393-398 (1954).
14. Hussey, G. *Scient. Prog., Oxford* **65**, 185-208 (1978).
15. Hussey, G. & Hephner, A. *Annls Bot.* **42**, 477-479 (1978).
16. Fosberg, R.A. & Smith, R.R. in *Hybridization of Crop Plants* (eds Fehr, W.R. & Hadley, H.H.) 65-81 (American Society of Agronomy and Crop Science Society of America, Madison, 1980).
17. Smith, G.A. in *Hybridization of Crop Plants* (eds Fehr, W.R. & Hadley, H.H.) 601-616 (American Society of Agronomy and Crop Science Society of America, Madison, 1980).
18. Rogers, H.H. *New Varieties of Cocoa in West Africa*, 76-79 (Report of the Cocoa Conference held at Grosvenor House, London 13th to 15th September 1955).
19. Burton, G.W. *7th International Grassland Congress Proceedings*, 439-448 (Palmerston North, New Zealand, 1956).
20. Glendinning, D.R. & Edwards, D.F. *Report of the 1961 Cocoa Conference*, 137-140 (1962).
21. McWilliam, J.R. *J. agric. Res.* **13**, 1-9 (1962).
22. Toxopeus, H. *Cocoa Growers' Bull.* **12**, 14-16 (1969).
23. Edwards, D.F. *Cocoa Growers' Bull.* **13**, 14-19 (1969).
24. Wood, G.A.R. *Cocoa 3rd edn* (Longmans, London, 1975).
25. Watts, L.E. *Flower and Vegetable Plant Breeding* (Grower Books, London, 1980).

NEWS AND VIEWS

Man the exterminator

from Jared M. Diamond

THE spread of European populations over the globe that followed the discovery of America soon led to the extinction of many species and local populations, including populations of man himself. Among the best known cases are those of the Dodo, Passenger Pigeon, Steller's Sea Cow, Tasmanian Wolf and Tasmanian humans. The colonizing Europeans often encountered local hunting groups who obeyed taboos that effectively promoted conservation of the wild life on which they depended. So arose the legend that man the hunter-gatherer is prudent, and only civilized man is an exterminator. Accumulating evidence, much of it recent or controversial, suggests that the truth may be very different.

Hunter-gatherers and neolithic agriculturalists seem now to have played the part of exterminator whenever they occupied new areas of the globe. That the areas occupied by man in the past 50 millennia have suffered extinctions, especially of large animals, is undisputed. What remains controversial is which extinctions preceded and which followed man's arrival and how the latter extinctions took place.

Many other waves of human migration preceded that of the Europeans. Only in Africa and Eurasia is man's antiquity measured in millions or hundreds of thousands of years. Present estimates suggest humans reached Australia and New Guinea 50,000 yr BP, North and South America 20,000 yr BP, the West Indies 5,000 yr BP, outer Melanesia (New Caledonia and Fiji) 3–4,000 yr BP, and Madagascar, New Zealand and Hawaii 1–1,500 yr BP.

The best understood of the early extinction waves occurred on New Zealand. In the 200 years since Europeans arrived, eight bird species are known to have become extinct. The discovery of moa bones, however, testifies to earlier extinctions. Until about 25 years ago, it was generally thought that moas had been dying out naturally^{1,2}, and that few remained when Polynesian colonists arrived around AD 1000. Excavations and radiocarbon dating have now shown that these colonists encountered at least 30 bird species, including

the 13 moas and 9 other flightless species, that had vanished by the time Europeans began surveying the avifauna in the 1800s^{3–8}. The moas disappeared in a north-to-south wave ending at the south of New Zealand's South Island in the 1600s⁹. In the same period, breeding fur seals became extinct on North Island (now confined to South Island), and large lizards, frogs and flightless insects became extinct on the New Zealand mainland (now confined to off-

and the extinctions are very recent and well documented. Yet it was long uncertain whether the New Zealand extinctions were even contemporary with human settlement, and it is not clear exactly how they took place, as is indeed often the case even for extinctions occurring under biologist's eyes today. No wonder then that the facts are disputed for earlier extinctions that are less accessible archaeologically, elsewhere in the world.

Around the time that Polynesians reached New Zealand, two other isolated islands with rich biotas were also first settled by man: Hawaii and Madagascar. When Europeans surveyed the Hawaiian avifauna in the nineteenth century, they found 49 land and freshwater bird species (two of them flightless), of which at least 15 are now extinct. This European-linked extinction wave is well known, but studies of subfossil bird bones reported in the current issue of *Science* show that it was preceded by an even larger Polynesian-linked extinction wave^{13,14}. At least 39 recent fossil species have been discovered that had disappeared before European arrival. Bones of most of these species have been found either in Polynesian archaeological sites or in deposits contemporary with these sites, proving that the extinctions followed the arrival of man. The lost prehistoric avifauna includes at least 7 species of goose, many of them flightless, plus two flightless ibis, seven flightless rails and at least fifteen subfossil species of the Hawaiian endemic tribe Drepanidini (the Hawaiian finches or 'honeycreepers'). The radiation of Hawaii's geese, as of New Zealand moas, presumably filled grazer niches in the absence of mammals other than bats.

Madagascar also had a remarkable subfossil fauna that disappeared before the Europeans arrived^{15–18}. It included about 14 large or giant lemurs, which together with about 19 surviving species constituted the richest sympatric primate community on Earth; 6–12 species of the giant flightless elephant birds (Aepyornithidae); 2 giant tortoises; and a hippopotamus, aardvark and large viverrid. The elephant birds and tortoises would have been the dominant terrestrial herbivores, equivalent to the moas and Hawaiian geese. Humans

IMAGE
UNAVAILABLE
FOR
COPYRIGHT
REASONS

shore islands).

Graphic witness to a role of overhunting in at least some of these extinctions are archaeological sites with stone ovens and hundreds of butchered moa skeletons⁵. But overhunting could hardly explain the extinctions of frogs and flightless insects. Even for the moas, it is hard to accept overhunting as the sole explanation on South Island, where human population densities were very low. Other causes, including the clearing of forest, predation by introduced Polynesian rats and introduced diseases, must also have contributed to the Polynesian-linked extinctions^{5–7,10–12}.

In New Zealand, both the arrival of man

From Tim Halliday *Vanishing Birds* (Sidgwick & Jackson, 1978).

Jared M. Diamond is Professor of Physiology at the University of California Medical School.

arrived around AD 500, and extinction of most of the megafauna apparently followed in the next six centuries.

For New Zealand, Hawaii and Madagascar the arrival of man was so recent, and (at least for New Zealand and Hawaii) the occurrence of extinct species in archaeological sites or levels so well documented, that the extinctions are obviously due to man rather than to late-Pleistocene climate changes. In the West Indies^{19,20}, the story becomes more complicated, because humans arrived earlier (about 5,000 yr ago) and fossil extinctions have yet to be dated. Fossil mammals included about nine ground sloths, numerous rodents up to the size of a bear, insectivores and a distinctive but enigmatic monkey. Carnivorous birds radiated in the absence of carnivorous mammals: a condor, giant eagle, barn owls of three different sizes (normal, giant and enormous), and an enormous and possibly flightless strigoid owl. Other extinct species include lizards and giant tortoises, as well as other birds. When each species went extinct is largely unknown, but there seems to have been at least three stages: late-Pleistocene extinctions due to moister climate and rising sea level and especially affecting species of dry habitats; a wave of extinctions when Amerindians arrived five millennia ago; and another, continuing wave that began with the arrival of Europeans in 1492.

On the North and South American continents a wave of extinctions peaking around 10,000–12,000 yr ago eliminated about three-quarters of the genera of large mammals (weight > 100 lb)^{15,21–26}. Although there is now no dispute that the extinctions occurred after man arrived, the question of whether the extinctions were due to man or to the late-Pleistocene climate changes remains controversial for several reasons. First, the New World differs from the previously discussed cases in that the arrival of a wave of human hunters around 12,000 yr BP (the Clovis hunters)²⁶ coincided with the rapid climate changes at the end of the Pleistocene. Second, most archaeologists believe that other humans had already reached the New World well before 12,000 yr BP, so that those who maintain the extinctions to be man-related have to postulate that the earlier colonists were either inefficient or indifferent big-game hunters. Third, clear associations of human artefacts with extinct fauna are very few, a fact which Mosimann and Martin²³ explained by a 'blitzkrieg model' of extinction — that is, hunters encountering native large animals decimate them so quickly that the period of coexistence between animals and man is brief, leaving few remains for archaeologists to distinguish from carcasses of animals that died naturally. This reconstruction seemed daring when first proposed, but it now appears to describe exactly the speed of extinction in New Zealand, Hawaii and Madagascar.

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

The alternative, climate-based interpretation of New World extinctions also faces difficulties. For instance, the Shasta ground sloth disappeared abruptly when Clovis hunters reached its range around 11,000 yr BP, yet radiocarbon-dated dung samples show it to have been common during the five centuries immediately preceding its extinction, and the plants on which the sloth fed still occur today at its fossil sites²⁴.

1. Oliver *Dominion Mus. Bull.* 15, 1 (1949).
2. Duff *Canterbury Mus. Bull.* 1 (1956).
3. Kinsky *Annotated Checklist of the Birds of New Zealand* (Reed, Wellington, 1970).
4. Kuschel *Biogeography and Ecology in New Zealand* (Junk, Hague, 1975).
5. Cassels *Wildl. mon. Mag.* 21, 31 (1979).
6. Cassels in *Quaternary Extinctions* (eds Martin & Klein) (University of Arizona Press, Tucson, in the press).
7. Mills & Williams in *The Status of Endangered Australian Wildlife* (ed. Tyler) (Royal Zoological Society of South Australia, Adelaide, 1979).
8. Anderson *Birds of a Feather* (British Archaeological Reports, Oxford, 1979).
9. McCulloch & Trotter *N.Z. archaeol. Ass. Newsletter* 18, 2 (1975).
10. Fleming *Notornis* 10, 113 (1962).
11. Williams *Notornis* 10, 15 (1962).
12. Bell *N.Z. Dept Internal Affairs Wildl. Publ.* No. 181 (1975).
13. Olson & Wetmore *Proc. biol. Soc. Wash.* 89, 247 (1976).
14. Olson & James *Science* 217, 633 (1982).
15. Martin & Wright *Pleistocene Extinctions* (Yale University Press, 1967).
16. Battistini & Richard-Vindard *Biogeography and Ecology in Madagascar* (Junk, Hague, 1972).
17. Tattersall *Primates of Madagascar* (Columbia University Press, New York, 1982).
18. Dewar in *Quaternary Extinctions* (eds Martin & Klein) (University of Arizona Press, Tucson, in the press).
19. Olson *Philadelphia Acad. nat. Sci. spec. Publ.* 13, 99 (1978).
20. Pregill & Olson *A. Rev. Ecol. Syst.* 12, 75 (1981).
21. Martin *Science* 179, 969 (1973).
22. Long & Martin *Science* 186, 638 (1974).
23. Mosimann & Martin *Am. Sci.* 68, 304 (1975).
24. Thomson *et al. Quat. Res.* 14, 360 (1980).
25. Martin & Neuner *Trans. Nebraska Acad. Sci.* 6, 117 (1978).
26. Haynes *Can. J. Anthropol.* 1, 115 (1980).
27. Jones *Archaeol. phys. anthropol. Oceania* 3, 186 (1968).
28. Gillespie *et al. Science* 200, 1044 (1978).
29. Martin & Murray *Abstr. 52nd ANZAAS Congr.* 25A, 5 (1982).
30. Green in *The Prehistory of Polynesia* (ed. Jennings) (Harvard University Press, 1975).
31. Olson *Smithsonian Contr. Paleobiol.* No. 23 (1975).
32. Bourne *Ibis* 109, 1 (1967).
33. Hadden *Birds of the North Solomons* (Wau Ecology Institute, 1981).
34. Greenway *Extinct and Vanishing Birds of the World* (Dover, New York, 1967).
35. Atkinson *Pac. Sci.* 31, 109 (1977).
36. Kepler *Auk* 84, 426 (1967).
37. Myers *Conversion of the Tropical Moist Forest* (National Academy of Sciences, Washington, 1980).

From Tim Halliday *Vanishing Birds* (Sidgwick & Jackson, 1978).

The other continent colonized by Pleistocene man, Australia, lost a bigger fraction of its megafauna than any other continent^{27–29}. Along with a giant snake (*Wonambi*), giant lizard (*Megalania*), giant tortoise (*Meiolania*) and giant flightless birds (*Genyornis*), Australia lost all its large mammals (>100 lb) except 4 kangaroos: 17 extinct genera and 46 extinct species of marsupials up to the size of the rhinoceros-sized herbivore *Diprotodon* and the lion-sized carnivore *Thylacoleo*, plus several species of giant kangaroos. As in the New World, the relative roles of climate and man in the extinctions are controversial. Evidence for association of megafaunal remains with human artefacts has yet to be found in Australia. Unlike the New World, Australia gained humans and lost megafauna long before the end of the Pleistocene. Humans arrived at least by 30,000 yr BP, and evidence for survival of megafauna much after that time is inconclusive.

These changes on land produced secondary effects at sea. It is likely that the populations of pelagic turtles³⁰ and birds³¹ that we see today are a tiny fraction of those existing before breeding colonies on oceanic islands were decimated by the first human colonists. Evidence includes abundant remains of subfossil petrels on Hawaii¹⁴ and in Polynesian middens on the Chathams³², and the recent discovery of breeding petrels in the remote interior of the Solomon island of Bougainville³³.

Evidence of man-linked faunal extinctions remains to be sought on many islands that have been little explored by palaeontologists but that were reached by man before the European expansion of the past 500 yr. Such islands include New Caledonia, New Britain, Fiji, the Solomons, Canaries, Comoros and Mediterranean islands. Of equal interest will be islands such as Lord Howe, the Galapagos and the remote islands of the Indian and Atlantic Oceans, that were not reached by man until the recent European expansion. Such islands will be controls that show the 'pure' effect of climate changes in the absence of man.

Man's hand fell especially on species that were large, flightless, or both. Large animals fared badly as the preferred prey of hunters, and because they tend to occur with low population densities and low reproductive potentials. Flightless birds fared badly as the easiest prey, whether to man or to his accompanying mammalian pests. For example, New Zealand lost its 19 largest birds; Madagascar, 14 of its 15 largest lemurs; Australia, 37 of its 41 largest marsupials; the New World, two-thirds of its largest mammals. An exception is Hawaii, which lost 15 small drepanidine finches along with large geese, crows and an eagle (see below). The Chathams lost all 6 of their flightless birds; Hawaii lost all but 2 of its 14–18 flightless birds to the Polynesians, and the remaining 2 to Europeans; and New Zealand lost 22 of its

29 flightless birds to Polynesians, and has lost or nearly lost 4 of the remaining 7 to Europeans. Dozens of other remote islands lost flightless birds to Europeans.

While the fact of human-related extinction waves is clearly established for some areas, at least six different mechanisms have been proposed, and their relative importance is debated. The most obvious is overhunting, the mechanism by which modern European hunters exterminated the Greak Auk and Steller's Sea Cow under the eyes of other Europeans. For example, the last two Great Auk were clubbed to death at Eldey Rock off Iceland on 3 June 1844. Animal populations that had never experienced man did not fear him, and this helps explain why extinction waves proceeded so quickly when man first reached remote islands and continents. Today one can still glimpse a surviving fragment of the world of 50,000 yr ago in New Guinea's isolated, completely uninhabited Foja Mountains. Here large marsupials are abundant and tame, in contrast to their rareness and shyness even in areas of New Guinea with only sparse populations of nomadic hunter-gatherers.

A second, related mechanism of man-linked extinction is the impact of predators introduced by man, such as rats, cats, dogs, weasels and mongooses, especially on nesting birds and ground-dwelling animals. Notorious examples include the extinctions of four birds, a bat and numerous invertebrates on Big South Cape Island off New Zealand within three years of the arrival of rats³; and a similar extinction wave after rats reached Lord Howe Island³⁴. The correlation between the spread of rats and the post-European extinctions of Hawaiian birds also suggests such an explanation³⁵. Even large animals can fall victim to man's pests: rats have been seen to kill adult albatross on the nest³⁶, and young of Galapagos Giant Tortoises are destroyed by feral pigs unless reared in captivity to a pig-proof size. Numerous species of New Zealand birds, reptiles, frogs and insects have disappeared from the mainland and now survive only on rat-free offshore islands.

A third mechanism attested by innumerable examples is destruction of habitat, whether by fire, by clearing or draining for agriculture, or by introduced browsing animals. Clearing of Hawaii's lowland forest probably explains why the Polynesian-linked extinction wave carried off at least 15 small songbirds not worth hunting, as well as 'game' birds such as geese¹⁴. Deforestation by fire and introduced animals eliminated much of the habitat on which Madagascar's megafauna depended¹⁸. By the time that the first European biologists arrived, Easter Island had no land birds at all, and St Helena only one species: their original complement disappeared with the cutting of the forests by Polynesians and Europeans respectively, before the biologists came.

Three other suggested mechanisms of

extinction involve introduced competitors, disease and social disruption. While man has exterminated native species, he has also introduced exotics that may compete with surviving natives. Disease is the primary mechanism by which colonizing Europeans have exterminated or reduced many non-European human groups they encountered in the last 500 yr. (Disease in the form of avian malaria may also have eliminated native Hawaiian forest birds from surviving forest below the malaria ceiling of 3,000 ft.) Finally, breeding in colonial species depends on social stimulation, so that a remnant population with its social structure disrupted may be incapable of recovery. This is probably why surviving Passenger Pigeons³⁴, and surviving human Tasmanians, dwindled to extinction after the hunting that had decimated their numbers ended.

What can past history teach us about future risks? We risk more of the same, in different places: overharvesting of the

cichlid fish fauna of East African lakes; arrival of rats at still rat-free Rennell Island, which has 5 endemic bird species and 17 endemic subspecies; and continuing destruction of Himalayan hill forest. But the major extinction wave of the near future will surely be through the swiftly advancing destruction of the continental tropical rainforests, the most species-rich habitats on Earth. While extinction waves cost New Zealand, Hawaii and Madagascar a considerable fraction of their biotas, these biotas were tiny compared with the biotas of Amazonia and tropical south-east Asia. The rainforests are already gone in much of the mountains of Colombia, and in coastal south-east Brazil. At present rates of destruction they will be virtually gone in lowland Malaysia, west Africa, Cambodia, Sri Lanka, Guatemala and Chiapas by the year 2000. Man the exterminator risks eliminating more species in the next few decades than in all his previous history. □

Twinkle, twinkle quasi star . . .

from Marek Abramowicz

WAS our Galaxy a quasar? Could it become a quasar in the future? In a recent issue of *Monthly Notes of the Royal Astronomical Society* (200, 247; 1982), Bailey argues that the Galaxy has indeed undergone at least one event of quasar-like activity in the recent past and may do so again.

Short, active periods recur in most galaxies. In Bailey's picture of galactic evolution, the mass lost by stars slowly settles in a central rotating disk. This continues quietly until the disk density exceeds M/r^3 , where M is the mass of the galactic nucleus and r is the disk radius. At this point the self-gravity of the disk becomes important and local instabilities, similar to the Jeans instability, develop. According to Paczyński¹, if the optical depth is so large that the cooling time scale is much longer than the rotational period, the instabilities will not produce condensations but will increase the velocity dispersion in the gas. Without condensations no star formation is possible. Velocity dispersion produces turbulence, thereby enhancing viscous stresses, which drive accretion by transporting angular momentum outwards and producing heat.

Bailey modifies this idea: he assumes that condensations will form and a sequential star formation occur. He suggests that this will produce turbulent motions in the gas which in turn increase viscosity. As in Paczyński's model, the

rapid viscous evolution of the disk fuels the activity of the nucleus by accretion of gas onto a supermassive central black hole. Alternatively, in the absence of such a hole, a massive central cloud may form which eventually collapses to form a black hole or spinar.

The evolution goes through the sequence of long, quiet phases interrupted by short periods of high activity; the quasars twinkle, very much like cataclysmic variables in our Galaxy. But does this idea agree with what we know about quasars?

Despite the substantial progress in our understanding of the nature of active galactic nuclei, a few fundamental problems have not yet been solved: (1) The supply problem — how can a galaxy supply enough matter to fuel the activity of its nucleus? (2) The efficiency problem — what physical mechanism provides a high enough efficiency of accretion? (Efficiency is the number of ergs of energy radiated per gramme of matter swallowed by the nucleus.) (3) The verification problem — how can observations confirm or rule out the different theoretical models (black hole/spinar)?

The supply problem

The galaxy supplies fuel to its nucleus either through mass loss from stars, or by more exotic processes of star captures and disruptions which can be very effective near the central black hole. Some questions remain open — among them, how to compute the stellar capture and disruption rates, and how to determine the fate of the resulting debris. As pointed out by Rees² and others, the classical estimations of

Marek Abramowicz is in the Department of Astrophysics, University of Oxford, South Parks Road, Oxford OX1 3RQ.

Hills are over-optimistic because they do not allow for the fact that once all the stars whose orbits pass close to the central hole have been destroyed, further capture can occur only when dynamical interaction between the stars repopulates the orbits, that is, on a time scale longer than 10^{10} yr. This means that the rate of supply from the galaxy may be insufficient to fuel continuously the activity of the nucleus. Obviously, Bailey's scenario gives an opportunity to solve this problem: the disk deposits a large mass by accumulating a low income over a very long time — it can therefore afford short time losses at a rate much higher than the income.

The efficiency problem

Black holes, once formed, cannot be destroyed. Soltan³ uses this fundamental fact to discuss the masses of quasars and the efficiency of accretion. If quasar-like events are common and if they involve black holes, then there should be black holes in most galaxies. The mass of the hole must be smaller than that of its host galaxy. Since we know the masses of galaxies from observation, we can set an upper limit to the black hole masses, which is also the upper limit for the mass of the fuel swallowed by the hole during the active phases. Soltan also estimates (from quasar counts and bolometric luminosities) an absolute lower limit for the total energy produced by the quasars per unit volume. His estimate is independent of cosmological model and agrees with a similar estimate of Schmidt. The upper limit for mass, together with the lower limit for energy, give the lower limit for efficiency, which is close to 0.1. Such a high efficiency cannot be consistent with very thick, stationary accretion disks. However, there are strong arguments in favour of very thick disks as models of active galactic nuclei. Thick disks can generate huge, super-Eddington luminosities observed in some BL Lac objects and produce long, jet-like structures in which matter moves with apparently superluminal velocities. It was suggested (by myself, Paczynski, Blandford and probably by others) that non-stationary disks can radiate away most of their energy while being geometrically thick from time to time. Between these flare-ups most of the matter is accreted onto a black hole from $r = 3 r_G$. Such disks may be both efficient and thick. Their behaviour resembles the short, active, recurrent events advocated by Bailey.

The verification problem

Bailey shows that the observations of several nearby galaxies are consistent with his theory. However, in the case of our Galactic Centre, his theory agrees with both the black hole and the spinar models, thus leaving unanswered the question, raised by Lynden-Bell, is there a black hole in the centre of our Galaxy? Previously Bailey has argued⁴ that the Galactic Centre may contain a compact object (black hole?) with a mass greater than $10^6 M_\odot$. Other authors, particularly Ozernol²,

maintain that in this case the Galactic Centre would have a much higher luminosity than is observed. This controversy will probably last for some time to come, because its source lies mainly in genuine difficulties of theoretical interpretation and cannot be resolved by progress in technology.

Bailey discussed some indirect observational tests of his theory. A direct test is also possible. Preliminary, unpublished results of Sikora and Soltan, who have analysed statistically the observed ratio of the radio luminosity of extended sources to the bolometric luminosity of their central nuclei as a function of redshift, seem to indicate that the galactic nuclei do indeed twinkle. The intensity and frequency of twinkling falls off with time.

Some of Bailey's calculations employ astrophysical theories that are still likely to experience new developments and

substantial revisions. In addition, a few of his results are in conflict with those of other investigators. This may indicate that some details of Bailey's theory will need reconsideration. However, the main idea, that active galactic nuclei switch on and off, or that 'quasars twinkle', is, in my opinion, both very important and correct. Some elements of this idea have been discussed by other authors before Bailey. However, he presented it for the first time in a wide context of galactic evolution and endowed it with astrophysically realistic scenarios. The richness and complexity of Bailey's theory will guide many theoreticians and observers in their work on active galactic nuclei in years to come.

1. Paczyński, B. *Acta astr.* **28**, 91 (1978).
2. Rees, M.J. Preprint *The Compact Source at the Galactic Center* (1982).
3. Soltan, A. *Mon. Not. R. astr. Soc.* **200**, 115 (1982).
4. Bailey, M.E. *Mon. Not. R. astr. Soc.* **190**, 217 (1980).

Opioids at Cape Cod

from Brian M. Cox

A DECADE ago the presence of a specific receptor activated by opiate drugs was accepted, but those who sought direct measures of drug binding to the receptor were still struggling to reduce background binding to acceptable levels. Only a few visionaries had begun the search for the endogenous ligand, and it was always referred to in the singular, if mentioned at all. Now the occurrence of at least three sets of endogenous opioids derived from three different genes is clearly established and, as emerged at the recent International Narcotics Conference*, evidence of an extraordinary degree of complexity in endogenous opioid systems is rapidly accumulating.

Conference sessions on the endogenous opioids were dominated by rumours that Numa's group in Kyoto had characterized a common precursor for dynorphin and the neo-endorphins. Now these rumours have been confirmed and the precursor sequence recently published in *Nature* (298, 245). Much evidence presented at the conference was consistent with Numa's findings. Measurements of tissue levels of dynorphin and α -neo-endorphin revealed closely similar anatomic and subcellular distributions (B. Seizinger, Max-Planck-Institut, Munich; C. Molineaux, Uniformed Services University, Bethesda). Immunocytochemical studies by E. Weber (Stanford University), recently discussed in these pages (*News & Views* 298, 221), were strongly suggestive of co-localization of these peptides in the same neurones.

Brian M. Cox is Professor of Pharmacology at the Uniformed Services University of the Health Sciences, 4301 Jones Bridge Road, Bethesda, Maryland 20014.

However, processing of the dynorphin precursor appeared to be differentially regulated in different tissues; both Seizinger and Weber provided evidence for the presence of substantial amounts of the dynorphin fragment dynorphin₁₋₈ in hypothalamus. The distribution of dynorphin in brain differs from that of the enkephalins and β -endorphin, although there are several areas where dynorphin and enkephalin precursors are both present (S. Watson, University of Michigan). Several studies emphasized that in the magnocellular neurones of the hypothalamo-hypophyseal tract, dynorphin and vasopressin are localized in the same cells, while K. Voigt (University of Ulm) reported that enkephalins are present in the oxytocin-containing cells.

But what of the roles of the products of the three opioid peptide genes? There is now general agreement that opioid receptors can be classified into at least three types, designated μ , κ and δ . The enkephalins have long been known to have preferential affinity for δ receptors. Over the past year it has become apparent that dynorphin has selective affinity for κ receptors; at the meeting, evidence was presented that the other products of the dynorphin gene, α -neo-endorphin, dynorphin B and dynorphin₁₋₈, also show marked κ -receptor selectivity (W. Fischli, Addiction Research Foundation; A. McKnight, Unit for Research on Addictive Drugs, Aberdeen). Thus the dynorphin gene generates opioids with preferential κ

*The International Narcotics Research Conference held its 13th meeting in North Falmouth, Cape Cod, Massachusetts on 13-18 June. The proceedings will be published later this year by *Life Sciences*.

activity and the enkephalin gene agonists with preferential δ activity. β -Endorphin is, however, more promiscuous, interacting with comparable affinity at μ , κ and δ sites. There is no known endogenous opioid with selective affinity for the μ sites, although both β -endorphin and the enkephalins have good affinity and could well function as physiological ligands at μ receptors.

Whether these various receptor forms represent different molecular entities or result from conformational changes of a single receptor species is still unclear. C. Pert (NIMH) argued that μ - and δ -binding sites in the striatum were interconvertible, the conversion being regulated by local Na^+ and GTP concentrations. However, I. James (Addiction Research Foundation) was unable to demonstrate any re-equilibration of receptor types in isolated brain membranes after the selective inactivation of one or more classes by the alkylating opioid antagonist β -chlor-naltrexamine. Autoradiographic studies suggest that the distributions of the μ - and κ -receptor forms in brain are different. S. Snyder (Johns Hopkins University) reported the presence of a particularly prominent band of κ receptors in the deep layers of the rat cerebral cortex, which he postulated might be linked to the sedative properties of κ receptor agonists. F. Leslie (University of California, Irvine) found the neonatal rat brain contained predominantly μ and κ receptors, with few δ receptors demonstrable. In the adult rat, the proportion of κ receptors falls substantially, while the proportion of δ receptors increases to levels comparable with μ -receptor densities. The rat, however, is apparently atypical in the low proportion of κ receptors observed in the adult animal. Higher levels have been observed in other species, and Y. Itzhak (New York University) reported that relatively high levels of κ receptors were observed in most regions of human brain.

The analysis of the functions of the different receptor classes has been facilitated by the development of several new drugs and reagents. The alkylating agents with opioid receptor affinity, such as β -chlor-naltrexamine, initially developed by P. Portoghese (University of Minnesota), have been found useful by several groups. G. Pasternak (Memorial Sloan-Kettering Cancer Center) reported the use of another irreversible receptor ligand, naloxazone, in the discrimination of opioid receptor subtypes. However, all these irreversible ligands have proved to have lower affinity for δ receptors than other receptor forms. The development, therefore, of a peptide analogue with selective competitive antagonist activity at δ receptors, reported by J. Shaw and M. Turnbull (ICI), promises to be very useful in identifying functions mediated by this receptor class. An approach to the identification of the roles of specific opioids in physiological processes is the use of selective inhibitors of degradation. R.

Greenberg (Squibb) and R.E. Chipkin (Schering) reported that thiorphan, an inhibitor of enkephalinase, potentiated analgesia induced by restraint or foot-shock stress. Interpretation of these results was complicated, however, by the observation that thiorphan also potentiated the analgesic effects of peptides not subject to degradation by enkephalinase.

Some progress has been made in identifying functions regulated by specific receptor subtypes. P. Wood (Douglas Hospital, Quebec) demonstrated that analgesia induced by opiate injection into the periaqueductal gray region of the midbrain is mediated by activation of μ receptors, while κ receptor-selective opioids are relatively inactive. However, M. Piercey (Upjohn) reported that injection of either dynorphin or a κ receptor-specific drug, U50488H, into the spinal cord in mice produced pronounced analgesia. The analgesic action of dynorphin in spinal cord was also confirmed by J. Han (Beijing Medical College). A possible role for dynorphin in the physiological regulation at the spinal cord level of the perception of noxious stimuli is suggested by Han's demonstration that electroacupuncture analgesia was suppressed by the localized administration to the spinal cord of dynorphin antiserum. The presence of relatively high concentrations of dynorphin in the dorsal

horn of spinal cord has previously been reported.

The role of opioids in the nigrostriatal system has been much studied because it is known that there is a very close relationship between dopaminergic and opioid systems in this part of the brain. Wood reported that μ - and δ -receptor agonists increased the levels of dopamine metabolites in the striatum, while κ -selective agonists, such as ethylketocyclazocine, had little direct effect, but antagonized morphine action. M-F. Chesselet (College de France) showed that the selective δ ligand D-Ser²-Thr⁶-Leu-enkephalin, injected into the fluid perfusing the caudate nucleus, induced a localized release of dopamine. Selective μ agonists such as morphine produced a weaker effect which was followed by inhibition of release. Since the enkephalin analogue, but not morphine, enhanced the release of dopamine from striatal slices *in vitro*, it was suggested that δ agonists act directly, while μ agonists act indirectly, to promote the release of dopamine in the caudate. The induction of muscular rigidity by morphine in the rat, probably by an inhibitory action of γ -aminobutyric acid neurones in the zona reticulata of the substantia nigra, was discussed by U. Havemann (Max-Planck-Institut, Göttingen). Thus in this, as in so many systems, several component structures seem to be influenced by one or other form of opioid. □

Do antigen-presenting cells distinguish self from non-self?

from Martin C. Raff

THERE is substantial evidence suggesting that helper T cells recognize a foreign antigen only after it has been taken up by specialized antigen-presenting cells (APCs) and that the antigen is recognized on the surface of the APC in association with class II (Ia) MHC glycoproteins¹. But this view of antigen presentation is fundamentally incomplete in that it does not provide answers to two central questions: how do APCs recognize antigens, and do they distinguish foreign molecules from self molecules?

If APCs do not discriminate between self and non-self, why are their recognition and processing pathways not saturated by the high concentration of self macromolecules in the extracellular fluid? If they do distinguish self from non-self, how do they do so? How can they pick out a molecule of tetanus toxin from a sea of self proteins? One possibility is that they use a non-immunological recognition mechanism akin to those used by certain invertebrate

cells for distinguishing self from non-self: although the molecular mechanisms involved are still poorly understood, it is clear that such discrimination operates when, for example, yeast cells mate², or a mixture of cells dissociated from two different species of sponges reagggregates in a species-specific way³, or invertebrate haematocytes react against foreign particles, cells or organisms⁴. However, the discrimination required of APCs, were they to distinguish the many thousands of different self molecules from the many millions of different foreign antigens that are capable of stimulating an immune response, would be far more demanding than that required of invertebrate cells. Teleologically, therefore, it seems more likely that APCs would take advantage of the sophisticated antigen recognition mechanisms evolved by the vertebrate immune system. For example, APCs might only efficiently bind and/or process molecules that are complexed to antibodies secreted by T cell-independent B cells. However, some APCs, such as splenic dendritic cells, apparently do not have Fc receptors with which to bind antigen-

Martin C. Raff is in the Department of Zoology, University College London, Gower Street, London WC1E 6BT.

antibody complexes⁵, and B cell-deficient animals seem to be able to make helper T cell-dependent immune responses⁶. Therefore, if APCs require lymphocyte help to recognize foreign antigens, this help might be in the form of antigen-specific T cell products (which must themselves be made without the aid of APCs). The recent report that suppressor T cells can inhibit the function of APCs⁷ is consistent with the view that communication between T cells and APCs is bidirectional.

The idea that antigen binding and/or processing by APCs is lymphocyte-dependent would be untenable if pure APCs (especially antigen-presenting cell lines) were found to be as efficiently 'loaded' with antigen in the absence of lymphocytes as in their presence. It would also be unlikely if the recent evidence (cited but not documented in ref. 8; and G. Sunshine, M. Cyrus and G. Winchester,

unpublished results) that APCs do not discriminate between self and non-self forms of the F serum glycoprotein is confirmed for other molecules. In either case, an important question in antigen presentation, which has so far been ignored, would remain unanswered: why are APCs not saturated by self molecules?

1. Rosenthal, A. S. *Immun. Rev.* **40**, 136 (1978).
2. Manney, T. R. & Meade, J. H. in *Microbial Interactions, Vol. 3. Receptors and Recognition Series B* (ed. Reissig, J. L.) 281 (Chapman & Hall, London, 1977).
3. Wilson, H. V. *J. exp. Zool.* **5**, 245 (1907).
4. Salt, G. *Parasitology* **53**, 527 (1963).
5. Nussenzweig, M. C. & Steinmann, R. M. *Immun. Today* **3**, 65 (1982).
6. Maguire, H. C., Fanis, L. & Weidanz, W. *Immunology* **37**, 367 (1979).
7. Ptak, W. & Gershon, R. K. *Proc. natn. Acad. Sci. U.S.A.* **79**, 2645 (1982).
8. Zitron, A., Mitchison, N. A. & Sunshine, G. H. *Immunobiology of the Major Histocompatibility Complex*, 7th int. Convoc. Immun., Niagara Falls, 243 (Karger, Basel, 1981).

Iron-sulphur cluster interconversions and the activation of aconitase

from Richard Cammack

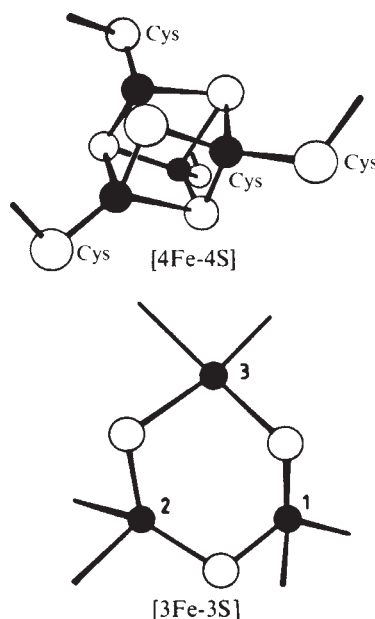
IRON-SULPHUR clusters in proteins may not always be such static entities as has been supposed. Recent experiments have shown that some of the newly discovered [3Fe-3S] clusters can be converted reversibly into [4Fe-4S] clusters. H. Beinert's group, working in Madison, Wisconsin, has proposed that an interconversion of this type is the key to the activation of the mitochondrial enzyme aconitase. It has long been known that aconitase, when isolated, is in an inactive form and can be activated by incubation with iron salts. Some authors have suggested that this iron atom is involved in the catalytic centre of the enzyme. Now it is recognized that aconitase contains an iron-sulphur cluster, of the recently discovered three-iron type (see *News and Views* **286**, 442; 1980). The additional iron appears to convert this to a [4Fe-4S] cluster which somehow is required for the activity of the enzyme.

Recent advances in the use of spectroscopic techniques, including ESR, magnetic circular dichroism (MCD), ⁵⁷Fe-Mössbauer and resonance Raman (RR) spectroscopy, have made it possible to observe the individual clusters in the iron-sulphur proteins. Though each technique provides different information about the type of cluster in a protein (paramagnetism, vibrational modes, and so on), they all provide ways of comparing an unknown iron-sulphur protein with the 'benchmark' proteins whose structures

have been determined by X-ray crystallography. A series of ferredoxins is now known which contain [4Fe-4S] and [3Fe-3S] clusters in various combinations. The advantage of studying these instead of proteins such as aconitase is that their relatively low molecular weights means that there is relatively more iron to observe.

The [4Fe-4S] cluster can be visualized as a cube, with iron and sulphide atoms at alternate corners. The [3Fe-3S] cluster is more like a flat hexagon of alternate iron

Atomic arrangements in the four-iron and three-iron clusters. Iron atoms are the solid circles.



and sulphur atoms¹. In each cluster, the iron atoms are also bound to ligands from the protein; in the [4Fe-4S] cluster, they are four sulphur atoms of the amino acid cysteine; the [3Fe-3S] cluster requires six ligands, as in ferredoxin of *Azotobacter vinelandii* are provided by five cysteine sulphurs and possibly a water or glutamate oxygen atom¹. Cluster interconversion therefore requires a considerable rearrangement of the protein structure.

The function of the clusters in most iron-sulphur proteins is to transfer electrons. (This cannot, however, be the case in aconitase, which is not an oxidoreductase.) A major class of iron-sulphur proteins has clusters that undergo oxidation-reduction between the [Fe-4S]²⁺ and [Fe-4S]¹⁺ oxidation levels. When such proteins are isolated, they often contain small and variable amounts of material which gives an ESR signal in the oxidized state of $g = 2.01$. The amount of this material can be increased by treatment with the artificial oxidizing agent ferricyanide. At first this was thought to be due to conversion to the [4Fe-4S]³⁺ oxidation level; however, if this occurs, the cluster is probably unstable and undergoes further modification, because the process cannot be reversed by simple reducing agents. The material which results from this modification, and which gives the $g = 2.01$ signal, has recently been shown by MCD², RR³ and Mössbauer⁴ spectroscopy to contain [3Fe-3S] clusters. Evidently oxidative treatment results in the loss of an iron atom from the [4Fe-4S] cluster, and rearrangement.

The change was referred to as 'oxidative damage'³, implying that the [3Fe-3S] clusters found are artefacts caused by the exposure of [4Fe-4S] proteins to oxygen during isolation. The question then arises as to whether any of the [3Fe-3S] clusters observed in isolated proteins are of physiological significance.

A test case is provided by the ferredoxins isolated from *Desulfovibrio gigas* by J. Le Gall (University of Marseille), J. J. G. Moura, I. Moura and A.V. Xavier (University of Lisbon). Although designated ferredoxins I and II, these are in fact two different assemblages of the same polypeptide of molecular weight 6,000. Fd I is a trimeric unit in which each protein subunit contains a [4Fe-4S] cluster; Fd II is a tetrameric unit in which most of the subunits contains [3Fe-3S] clusters. Both ferredoxins are obtained in substantial quantities when the bacteria are extracted. They have different redox potentials and different activities with the enzymatic systems of *Desulfovibrio*. Fd I is active in the phosphoroclastic cleavage of pyruvate while Fd II is more active in sulphite reduction⁵. Various treatments have been found to interconvert the ferredoxins. Incubation of Fd II with cell extracts and pyruvate (reducing conditions) causes the appearance of Fd I, as seen by ESR. Oxidation of Fd I with ferricyanide

Richard Cammack is a lecturer in the Department of Plant Sciences, King's College London, 68 Half Moon Lane, London SE24 9JF.

converts it into something resembling Fd II. The protein has six cysteines and thus can readily accommodate a [3Fe-3S] cluster. The finely tuned balance between the cluster types was shown by reconstitution experiments in which the iron and sulphide were removed entirely and then replaced⁶. With an excess of iron present (5 per protein molecule), only [4Fe-4S] clusters were produced. With limiting iron (3 per protein molecule), up to 50 per cent of the clusters formed were of the [3Fe-3S] type.

Moura *et al.*⁶ have developed a procedure to add the extra iron atom to Fd II and produce [4Fe-4S] clusters. This provides a unique opportunity to examine a specific iron atom in a [4Fe-4S] cluster. Using Mössbauer spectroscopy one can observe the isotope ⁵⁷Fe (which has an abundance of only 2.2 per cent in natural iron), so that by adding either enriched ⁵⁷Fe or ⁵⁶Fe to Fd II it is possible to obtain spectra from either the added iron atom, or the average of the other three. The studies were carried out by E. Münck and co-workers in Minnesota.

The spectra show that the added iron atom probably goes into a specific site in the [4Fe-4S] cluster, though the possibility of two equivalent subsites cannot be excluded⁷.

The inequivalence between the iron atoms in this type of cluster is not understood at present. For example, in the reduced cluster, when the valence state corresponds to one Fe III and three Fe II atoms, the iron atoms as seen by Mössbauer spectroscopy appear to be present in two equivalent pairs⁸. A theoretical understanding of these clusters

can now be assisted by the determination of the electron distribution and hyperfine field at the individual iron atoms.

Similar experiments have been carried out on the iron-sulphur cluster in aconitase⁹. The enzyme is activated by reductive treatments in which an iron atom is introduced into a particular site in the [4Fe-4S] cluster. The incorporated iron atom is relatively loosely bound, since added ⁵⁶Fe will exchange in a few minutes with ⁵⁷Fe-activated enzyme. By contrast, in *D. gigas* Fd I, the incorporated Fe will not readily exchange. Oxidative treatments, which reconvert the cluster to the [3Fe-3S] form, inactivate the protein, but it is still not clear why.

It has not yet been demonstrated that this interconversion occurs *in vivo*, apart from an early report¹⁰ that the ESR signal at $g = 2.01$ resembling that of the aconitase [3Fe-3S] cluster can be observed in mitochondria. If it does, it would represent a totally new form of covalent modification of enzyme activity. It may not be unique; other enzymes, notably some hydrogenases, exist in inactive forms and show $g = 2.01$ ESR signals. □

1. Ghosh, D., Furey, W. Jr, O'Donnell, S. & Stout, C.D. *J. biol. Chem.* **256**, 4185 (1981).
2. Thomson, A.J. *et al. Biochim. biophys. Acta* **637**, 423 (1981).
3. Johnson, M.K., Spiro, T.G. & Mortenson, L.E. *J. biol. Chem.* **257**, 2447 (1982).
4. Bell, S.H. *et al. FEBS Lett.* **142**, 142 (1982).
5. Moura, J.J.G., Xavier, A.V., Hatchikian, E.C. & Le Gall, J. *FEBS Lett.* **89**, 177 (1978).
6. Moura, J.J.G. *et al. J. biol. Chem.* (in the press).
7. Kent, T.A. *et al. FEBS Lett.* **138** (1982).
8. Middleton, P., Dickson, D.P.E., Jonson, C.E. & Rush, J.D. *Eur. J. Biochem.* **88**, 135 (1978).
9. Kent, T.A. *et al. Proc. natn. Acad. Sci. U.S.A.* **79**, 1096 (1982).
10. Ohnishi, T., Ingledew, W.J. & Shiraishi, S. *Biochem. biophys. Res. Commun.* **63**, 694 (1975).

detailed phylogenetic trees were constructed. Now the first results of DNA sequence analysis of mitochondrial (mt) DNA are becoming available. In 900 base pairs (bp) of DNA sequence around the D-loop region of human mtDNA from seven individuals, nucleotide substitutions were found at 45 sites and DNA insertions at three positions (C. Aquadro, National Institute of Environmental Health Sciences, North Carolina).

Nucleotide substitutions were almost always transitions (that is, A→G or T→C changes), transversions accounted for only four per cent of the total. A phylogenetic tree, constructed by a maximum parsimony procedure that minimizes the number of changes between each sequence in the finished tree, reveals that convergent or parallel substitutions have occurred in at least five positions. In addition, the distribution of substitutions along the sequence was highly non-random because multiple substitutions occur at certain sites. These results raise doubts about the estimation of genetic variability from DNA sequence data that depend on pairwise comparisons, for these would greatly underestimate the sequence change inferred by the phylogenetic approach. Perhaps more worrying is that a restriction enzyme analysis would produce quite incorrect trees, because of the localized occurrence of base substitution.

Turning to nuclear genes, more is known of the evolution of the mammalian β -globin gene family than any other system. In man, these genes are arranged in the order:

$$5' \psi\beta_2, \epsilon, \gamma, \gamma, \psi\beta_1, \delta, \beta 3'$$

Sequence homology between the δ - and β -globin genes in man suggested that they arose by duplication about 40 million years ago. However, DNA sequence analysis shows that the 3' portion of the mouse $\beta H2$ gene has nearly 80 per cent homology to the human δ gene but only 58 per cent to the human β gene (M. Edgell, University of North Carolina). In addition, part of the single pseudogene in the lemur β -globin cluster appears to be a δ -globin sequence that has undergone gene conversion with an embryonic gene (A. Jeffreys, University of Leicester). These results strongly suggest that the β/δ -globin duplication predates the mammalian radiation of 80 million years ago. Where gene conversion can occur, overall sequence homology cannot be used to estimate the times of gene duplication events.

A large body of data collected on variants in the human β -globin gene cluster has revealed some remarkable aspects of the relationship between linkage disequilibrium and recombination distance (H.K. Kazazian, Johns Hopkins University). Thirteen polymorphic restriction site markers have been studied in several hundred individuals. Seven are located in a

Andrew Leigh Brown is at the Imperial Cancer Research Fund Mill Hill Laboratories, Burton-hole Lane, London NW7 1AD.

Evolution and molecular biology

from Andrew Leigh Brown

DESPITE the flush of publications commemorating the centenary of Darwin's death, those outside the field could be forgiven for believing that evolutionary biology has lost its way. While the horizons of the rest of biology are being dramatically extended by new technologies, evolutionary biology seems to be becoming more and more introspective. A recurrent discussion of 'tempo and mode' in evolution has reached such prominence that 'punctuated equilibria' appear to constitute the only issue of importance. In fact, this is far from being the case.

A more fundamental problem raised by developments in molecular biology has so far passed almost unnoticed: the concepts and language of population genetics, in

which the behaviour of genes in populations is described in the terms of classical genetics, are no longer adequate. The very words 'locus' and 'allele' have become almost meaningless. A new phraseology is required which can accommodate data on both restriction site polymorphisms and nucleotide substitutions, as well as more radical types of genetic variation: insertion polymorphism, 'concerted evolution' and transposition of repeated elements. Some of the first steps in erecting this new framework could be seen at two recent meetings* intended to foster understanding of the relationship between molecular and evolutionary biology.

The mitochondrial genome was used for the first essays in the molecular study of evolution because of its small size and ease of isolation. Population structure was analysed by restriction site variation and

*The Annual Meeting of the Society for the Study of Evolution and the American Society of Naturalists was held at Stony Brook, New York from 21 to 24 June 1982. The London Darwin Meeting was held on 5 July 1982 at University College London.

32-kilobase (kb) region on the 5' side of the δ -globin gene which includes the γ , ϵ and $\psi\beta_1$ genes. Six sites are found in a separate 20kb block across and to the 3' side of the β gene. Theoretically, the variants at the seven restriction sites might be found in over one hundred different combinations in the 5' region. Remarkably, only three combinations are common, all of which are distinguished from each other at at least two sites. Similarly, in the 3' block, only three more combinations are found. This high degree of linkage association would be unsurprising in view of the short distance separating the markers, but for the observation that across an 11 kb region spanning the δ gene, random association between 5' and 3' haplotypes is found. However, this recombination hot spot is unlikely to be a function of the δ -globin gene itself, as the flanking blocks include all the other genes in this cluster.

In surveys of the β -globin region, few DNA insertion variants have been found. In contrast, studies of restriction site polymorphism at two genes in *Drosophila* indicate that insertion/deletion events are an important source of variation. In a 12 kb region around the alcohol dehydrogenase (*Adh*) gene in *D. melanogaster*, five insertion/deletion variants were discovered in 18 lines from natural populations (C.H. Langley, NIEHS, North Carolina). However, the commonest restriction map in *D. melanogaster* showed little divergence that could be ascribed to DNA insertion from those obtained for two other species, *D. simulans* and *D. mauritiana*. Similar results are seen for the 87A heat-shock locus in *Drosophila* (A. Leigh Brown, Imperial Cancer Research Fund). It is not yet known whether these other *Drosophila* species also show insertional variation but it could be that 'the spaces between the genes' are important in as yet undefined ways such that insertions or deletions are generally screened out by natural selection. The strong conservation of the distances between genes in the β -globin cluster of higher primates has prompted similar speculations (Jeffreys).

The *Adh* gene in *D. melanogaster* is also the first nuclear gene to be used for a systematic survey of variation in natural populations using DNA sequencing (M. Kreitman, Harvard University). Of 1,750 bp which have been sequenced in 11 isogenic strains, 36 nucleotide positions are variable. Unlike mtDNA, transitions and transversions are equally frequent. No amino acid substitutions were detected apart from the Thr-Lys substitution which causes the well known 'fast-slow' electrophoretic polymorphism. If it is assumed that the only substitutions permitted are silent ones which do not result in an amino acid change, the frequency of polymorphic nucleotide sites relative to the number of silent sites in exons is very similar to the total frequency found in the introns. A much lower frequency was found in the 5'-non-coding regions of the transcripts,

suggesting that some constraints on nucleotide substitution may be acting upstream of the protein-coding region. However, these apparently do not extend to the transcription start point, around which substitutions were quite frequent.

Finally, population genetics theory is now being applied to the behaviour of transposable elements in eukaryotic populations. A simple mathematical model has been developed that gives the expected frequencies of occupation of a particular chromosomal site by a mobile element in a natural population under the action of random genetic drift (J. Brookfield and C.H. Langley, NIEHS). Testing this model with data on the sites of hybridization to *Drosophila* polytene chromosomes of three mobile elements, *copia*, 412 and 297, Brookfield and Langley showed that the data give a good fit for a high transposition rate. The result may be unexpected given that some mutations due to *copia* insertions (for example w^a) are very stable. It was also observed that the frequency of co-occupancy of a site by more than one type of transposable element was much higher than expected. This suggests either that insertion is not random and occurs after strand breakage or some similar event, or that the total number of occupable sites is restricted.

The P factor is a transposable element whose movement is responsible for a syndrome of germ-line aberrations in *D. melanogaster* known as hybrid dysgenesis. These aberrations are only found when wild males are crossed to females from long-established laboratory strains. Wild-type strains from the USA almost without exception carry this factor (P strains) while old-established laboratory strains do not (M strains). It had been thought that this loss was a consequence of inbreeding but new data on the relative frequency of such strains in collections from Europe and the Far East give a very different picture (M. Kidwell, Brown University). Here 'Q' strains, which are compatible with both P and M strains but contain many copies of

the P element, are common, while M strains are of intermediate frequency and P strains are rare. In addition, some wild M strains are now known to contain some copies of the P element. The P element is known to be highly labile and it is not known whether any of the elements in these strains are functional. This complex pattern suggests that the element is itself undergoing rapid evolutionary change and possibly indicates an invasive rather than a symbiotic history.

Speculation that transposable elements may be involved in speciation has flourished following the discovery of their role in hybrid dysgenesis. If, however, the germ-line incompatibility is a consequence of a transient process of invasion, such involvement is much more difficult to envisage. It is hardly reasonable to propose a new invasion for each speciation event. Similarly, no clear pattern can yet be discerned on the general significance of insertional variation, which is apparently common in *Drosophila*, but is very rare in the human β -globin region. It seems more reasonable that the current difference between the mammalian and *Drosophila* data is due to the particular genes which have been studied in each organism to date [for example, the *dilute* coat colour mutation in the mouse is due to an insertion of retro-viral sequences (Jenkins, N. *et al.* *Nature* 293, 370; 1981) and insertional polymorphism has also been found in man (Wyman & White, *PNAS* 77, 6754; 1980)].

These studies are beginning to reveal some of the processes which take place in the genome during evolution. It should be nevertheless recognized that the pursuit of detail is not a satisfactory end in itself. However much is learnt about a single gene, if the knowledge cannot be generalized we will not have discovered much about evolution. The greatest interest in the new developments stems from the fact that the whole genome has been rendered accessible by these techniques, and that is something that evolutionary biologists have waited a long time for. □



100 YEARS AGO

THE two Swedish gunboats which conveyed the circumpolar observation party to Spitzbergen, have just returned to Tromsø. Capt. Palander states that it was impossible to approach Mossel Bay, he having made two attempts, on account of heavy pack-ice, and that the party had therefore settled on Cape Thordsten as their residence, where observations commenced on the 15th inst. He further reports that all the Norwegian fishermen he met complained of the unfavourable season

and the enormous quantity of ice this summer, no vessel having been able to get higher than Amsterdam Island, from where no opening could be seen by telescope in any direction.

The steamer *A.E. Nordenskjöld*, belonging to M. Sibirakoff, left Tromsø on the 18th inst. for the Jenisei. The vessel has on board a cargo of English merchandise, two steam launches, and an engine for the Siberian gold works. The vessel will attempt to save some of the cargo lost in the *Oscar Dickson* in Gydaviken, and return next year with a full cargo of tea, which will be brought from China across the Baikal sea to Kureika on the Jenisei.

The death is announced of Count Lutke, well known in connection with Russian Arctic exploration, especially in the Novaya Zemlay region.

From *Nature* 26, 447; August 31, 1882.

Detectors in astronomy

Optical astronomy

from Alec Boksenberg

It is now more than a century since the invention of the gelatin photographic emulsion, but the photographic process still has many deficiencies, including nonlinear response, restricted dynamic range, reciprocity failure, adjacency effects, and a detection efficiency of only a few per cent. It does, however, have the advantage over all other detection media when a huge number of picture elements is required. However, for most detection and recording applications in astronomy it has been superseded by more efficient methods.

Direct interaction of photons with the recording medium occurs in photographic emulsion, electronically-scanned silicon devices such as the charge-coupled-device (CCD), and electron-beam-scanned television camera tubes of the simple vidicon type. In all other devices a photocathode is used to produce an intermediate electron image, which then is accelerated to relatively high energy to enable the signal to be enhanced by some gain process before recording.

Electronographic cameras

In 1934 Kiepenheuer proposed that an efficient detector could be made by accelerating and focusing electrons from a photocathode onto a photographic emulsion. This would give a considerable gain in sensitivity over a simple photographic emulsion because every high-energy electron entering the emulsion produces developable grains and there is no reciprocity failure. The overall efficiency should therefore depend mainly on the efficiency of the photocathode. Modifications to the technique, particularly by McGee, who in his 'Spectracon' used a mica barrier membrane to protect the photocathode and also allow the emulsion to be used in air (fine-grained nuclear emulsions are used in modern devices), and the later variant by McMullan, have resulted in more convenient operation.

Photocathodes are now fundamental in certain television camera tubes and in image intensifiers. The most commonly used photocathode in optical astronomy, trialkali, has an efficiency of about 20 per cent in the blue falling to about 1 per cent at 8,000 Å. Recently a new class of 'negative electron affinity' photoemitting materials,

such as GaAs(Cs), have been discovered which have a high efficiency in the red and near infrared.

Integrating television camera tubes

Conventional television camera tubes can in principle be used to detect and record astronomical images, however, television broadcast cameras are severely limited by noise (from the video pre-amplifier) at low light levels. It is therefore important to use a camera tube with high internal gain and the capability to integrate a weak signal over a long period before readout. Two tubes having these properties are the SEC and SIT vidicons, both originally developed for the military market as night vision devices. A photocathode generates photoelectrons which are accelerated to a few keV in the image section and focused at the target. Electron multiplication takes place in the target and, at the end of the exposure, the amplified electron image is read by a scanning electron beam in the readout section. The SIT target has an electron amplification up to 2,000 but requires strong cooling (to -60°C) in order to reduce the dark current sufficiently to permit long integration times. The SEC target has lower gain (~ 100) but can integrate and store charge images indefinitely at room temperature without degradation, but it has the disadvantage of lower storage capacity and produces 'grainier' images. Notwithstanding, the SEC tube has been used successfully in several astronomy programmes, and is the prime detector on the International Ultraviolet Explorer observatory satellite where its operational simplicity is fully exploited.

Image intensifiers

Image intensifiers come in several forms, all using a photocathode for the initial conversion of photons into an electron image. They have arbitrary input and output spectral characteristics, for example a device may accept a UV image but convert it to an enhanced optical image. In the simplest case electrons from the photocathode are accelerated onto a phosphor screen to produce an enhanced optical image. A blue-light gain of 50 is typical for a simple device of this kind, but several such stages can be arranged in cascade to give a very high overall gain, a four-stage intensifier produced by EMI having a blue-light gain of about 10^7 .

The microchannel plate (MCP) can also

be employed as an image intensifier capable of extremely high gain. It consists of an array of fine, slightly conducting glass tubes having a secondary emission coefficient greater than unity. An accelerated electron entering one of the channels and striking the wall produces secondaries, which are accelerated by an internal electric field and in turn strike the wall, eventually producing an avalanche of electrons which emerges from the output of the channel.

The MCP can also be built into a very compact double proximity-focused structure using a conventional photocathode on its input window to give a

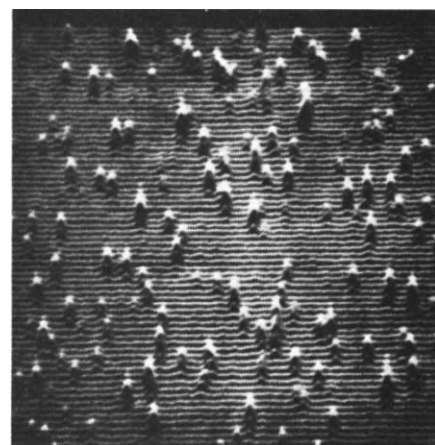


Fig. 1 Single photon events detected with a high gain image intensifier viewed by a television camera, as employed in the IPCS. This isometric display represents less than half of one per cent of the available television field.

'wafer' intensifier with high gain and no distortion. Although the electron gain of a MCP with straight channels is limited by the onset of positive ion feedback, several techniques have now been introduced to suppress this, the most satisfactory being to curve the channels, which can give an electron gain greater than 10^7 without difficulty. Additionally, the output pulse sizes at such high gains are confined to a sharply peaked distribution which substantially decreases the added noise in the image compared to the low gain performance. Although the output of an image intensifier may be recorded photographically, this combination still suffers from non-linearity and small

Alec Boksenberg is Director of the Royal Greenwich Observatory, Herstmonceux Castle, Sussex.

dynamic range. Coupling to a photo-electronic detector such as a television camera tube is far more satisfactory. This is exploited in the widely used technique of 'image photon counting' first developed by Boksenberg, who employed an EMI four-stage intensifier lens-coupled to a Philips 'Plumbicon' television camera tube.

Image photon counting systems (IPCS)

The basic idea of image photon counting is to amplify a faint image to a level where single photon events can be discerned above the readout noise of a television camera used to view the output screen of a high gain image intensifier (Fig. 1). The television camera acts as a spatially sensitive one-frame buffer store and passes the photon signals to a special pattern processor which records their central positions, and then passes these to the digital electronic memory of an on-line computer. Photon events registered in each television frame are then accumulated digitally. There is no limit to the length of exposure attainable because the system has no saturation capacity (unlike photographic emulsion or television camera targets). The technique shifts the emphasis from the detailed performance of the individual devices themselves to the real-time data processing of the fundamental signals delivered by the initial photon conversion layer in the system.

Recording only the central positions of each detected photon has two major advantages. Maximum efficiency is attained because each photon event is accorded equal statistical weighting, unlike the raw signals delivered by the television camera which are variously sized due to unequal gain processes associated with the coupled image intensifier. This would otherwise lead to uncertainty in the number of detected events. Second, a markedly sharper image is obtained because the image smearing inherent in the devices is removed in the processing of the photon events. Also, since all amplifier noise is rejected, there is no low-level threshold for very faint images, so images containing even a few photon events can be recorded accurately. Another major advantage of the system is the real-time display of the accumulating image on a monitor, which results in high operating efficiency in astronomy; considerable time may be saved by terminating exposures as soon as the required image signal-to-noise ratio has been achieved.

Two other systems also in routine use are the Robinson and Wampler image dissector scanner, and the intensified Reticon scanner developed by Shetman. These have more restricted channel capacity than a full IPCS but are excellent for many one-dimensional spectroscopic applications. Several more forms of image photon counting (or quasi-counting) detector are now being developed, employing various forms of intensification and readout, some resulting in very compact devices.

Solid state detectors

Various types of solid state imaging detectors based on silicon photodetectors have become available for astronomy in the past few years. Silicon is particularly efficient in the red and near-infrared compared to photoemissive surfaces, reaching about 80 per cent between 5,000 and 7,000 Å and extending beyond 11,000 Å. The various devices such as the Reticon and the charge-coupled device (CCD) are distinguished by the method of reading out the stored image. The detection efficiency is more or less limited by readout noise and build-up of charge background, and is highest for the CCD.

CCD devices all contain a two-dimensional array of sensitive elements (typically of dimension 15–30 µm), and arrays of up to 1,500 × 1,500 elements are under development. The device is essentially a silicon integrated circuit of MOS type and comprises an oxide covered silicon substrate upon which is formed an array of closely spaced electrodes. Each electrode is equivalent to the 'gate' of an MOS transistor. Signal information is carried in the form of an electric charge which is localized beneath the electrodes with the highest applied potentials. Charge accumulates in these potential wells and is released for conduction in the silicon by the action of the incident photons. 'Charge coupling' is the technique by which this signal charge can be transferred from beneath one electrode to the next. This is achieved by sequentially pulsing the

voltages on the electrodes between high and low levels. Charge can therefore be made to pass down an array of many electrodes with hardly any loss and very little addition of noise. By this means the entire charge image can be shifted out bodily, and entered into a line readout section which delivers the now sequentially arranged signals to an on-chip video amplifier. The noise in this amplifier represents a limit to the minimum detectable signal, but in current devices can be as low as six electrons per image element. Cosmic ray events represent another limitation as silicon is an efficient detector of high-energy charged particles. For use in astronomy the devices must be cooled to avoid the build-up of thermally-induced charge, which would otherwise mask the signal. Exposures of several hours can then be achieved without excessive interference due to dark current.

Future improvements in image detection are likely to come at least as much from the application of new electronic processing and data reduction methods as from the development of better devices. The CCD is taking an increasing role in astronomy, and in future all detector devices may be of this type. Vast arrays might be built up of CCD configurations and internal gain might be achieved by avalanche processes already implemented in discrete diode structures. The technique of image photon counting, with its superior low light level performance, may then be achieved entirely in solid state. □

Infrared astronomy

from Erick T. Young

SINCE the first detection of infrared (IR) radiation from a celestial source by Sir William Herschel¹ in 1800, the sensitivity of IR detectors has improved by many orders of magnitude. However, the true revolution in IR astronomy came in the 1960s with the advent of cooled semiconductor detectors. With these devices, the barrier to detecting ever fainter sources became the natural backgrounds from the atmosphere and telescope rather than the detector itself.

The astronomical IR spans the wavelength range from just below 1 µm to 1,000 µm. Although groundbased observations are limited to a few atmospheric windows of below 30 µm and at 350 and 450 µm, observations from balloons, aeroplanes, and spacecraft have now opened the entire range to study. Detection techniques vary from optical methods at the shorter wavelengths to quasi-radio methods in the submillimetre range. The choice of detector for a system depends on

cost, required sensitivity, wavelength coverage, backgrounds, cooling capabilities, and time resolution. The requirements of IR astronomy are quite distinct from most applications. Most importantly, astronomical sources are exceedingly faint, and consequently sensitivity is the most important detector parameter. To this end the detectors are usually operated cooled with restricted fields of view to keep the background noise to a minimum.

Sensitivity limits for detectors

The sensitivity of an IR detector is the ultimately limited by noise both in the detector and in the associated electronics. Thermal detectors are subject to phonon noise associated with the finite temperature of the sensing element. Photon detectors can have noise due to the random generation

Erick T. Young is at the Steward Observatory, University of Arizona, Tucson.

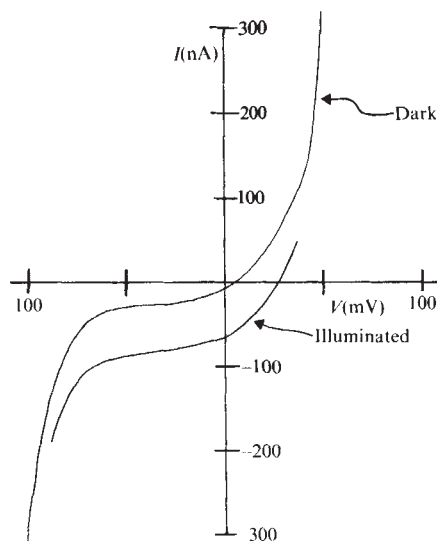


Fig.1 Current-voltage curves for an InSb photovoltaic detector operating at 77K. The two curves represent the case with background irradiation and the case without background irradiation.

and recombination of charge carriers, and all detectors suffer from Johnson noise due to the thermal motions of the charge carriers. Additionally, noise due to the fluctuations in the background photons seen by the detector can contribute significantly, and indeed are the most important source of noise with the current generation of IR detectors. The great promise of space experiments is the elimination of backgrounds due to the atmosphere and the telescope.

For an IR astronomer, the two most important parameters for detectors are the noise equivalent power (NEP) and the quantum efficiency η . The NEP gives the power incident on the detector which yields a unity signal-to-noise ratio in one second of integration time, and is therefore an indication of the required integration time for detecting an astronomical source at a given signal-to-noise ratio. The responsive quantum efficiency η is the ratio of electrons generated in the detector to incident photons. Of more use is the detective quantum efficiency η_D , first applied to IR detectors by Jones², $\eta_D = \text{NEP}_{\text{BL}}^2 / \text{NEP}^2$, where NEP_{BL} is the noise equivalent power of a perfect detector, limited only by the background photon fluctuations, and NEP is the actual noise equivalent power for those conditions. η_D is therefore a measure of how closely a detector approaches the theoretical limit. The detective quantum efficiency is especially useful when the photon background fluctuations are the dominant noise source. In the absence of spurious noise, η_D has the same value as η .

Thermal detectors

Herschel's first IR detector was a thermal detector, a thermometer. Thermal detectors utilize the temperature dependence of some physical parameter for the

detection of IR energy. The IR energy is converted into thermal energy of the lattice which is subsequently sensed. Currently, the most sensitive thermal detectors are liquid helium cooled germanium or silicon bolometers. With suitable doping, the impurity conductivity in silicon and germanium shows a very strong dependence on temperature³. By carefully mounting the semiconducting element so that the thermal conductance to the helium bath is small, minute amounts of IR energy will cause measurable changes in the conductivity of the device. Clearly the thermal fluctuations in the device can be a significant noise source, and it is highly advantageous to cool the detector. The gallium-doped germanium bolometer invented by Low⁴ was the first device to approach the theoretical sensitivity limits at liquid helium temperatures. In the background conditions where the noise due to photon fluctuations is greater than the thermal fluctuation noise, the Ga-doped Ge bolometer is a virtually ideal detector with almost unity quantum efficiency. This detector is widely used in astronomical observations, and its application to photometry has been described in detail by Low

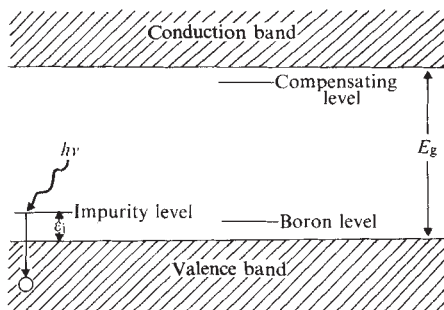


Fig.2 The band diagram of a typical extrinsic semiconductor with a shallow p-type dopant.

and Rieke⁵. Typically these devices are cooled with superfluid helium to below 1.5 K, and NEPs as low as $7 \times 10^{-15} \text{ W Hz}^{-1/2}$ have been realized. Because bolometers are thermal detectors, the wavelength limits to the response are set by the efficiency of coupling the radiative energy into the lattice. Using special absorbing materials, sensitive detectors into the millimetre wavelength range have been made. Current directions in research include developing bolometers for use at the much lower temperatures available with He III refrigerators and dilution refrigerators. NEPs better than $10^{-16} \text{ W Hz}^{-1/2}$ appear to be realizable.

Photovoltaic detectors

At wavelengths below $5 \mu\text{m}$, the most commonly used IR detector is the InSb photodiode, which has generally replaced the older PbS detector in high performance applications. In groundbased observing conditions, the InSb detector offers excel-

lent performance, with quantum efficiencies in excess of 60 per cent. In a photodiode, charge carriers are generated by photons in the depletion region and are swept out by the p-n junction voltage. Figure 1 shows the current-voltage curve for a high performance InSb photodiode. In the presence of photons, a current is generated across the junction that is proportional to the irradiance. These detectors became astronomically useful with the incorporation of the transimpedance amplifier (TIA)⁶. This amplifier acts as a current to voltage converter, and by maintaining a zero potential across the detector, noise contributions due to leakage currents are eliminated. The Johnson noise of the detector and noise in the associated electronics become the ultimate performance limitations. By cooling the detector to liquid helium temperatures and by using very low noise junction field effect transistors (J-FETs) in the amplifier, Rieke *et al.*⁷ have attained background-limited performance at wavelengths as short as $2 \mu\text{m}$. NEPs of $7 \times 10^{-17} \text{ W Hz}^{-1/2}$ have been reached, with operation limited to low frequencies.

Photoconductive detectors

When electron-hole pairs are created in a semiconductor by the absorption of photons, the conductivity of the material is temporarily increased until these charge carriers recombine. Application of a bias voltage allows the external detection of these carriers in an amplifier. Intrinsic photoconductivity can occur if the energy of the ionizing photon is greater than the bandgap of the material. As the intrinsic bandgaps of Si and Ge are 1.166 and 0.75 eV, respectively, these materials are not suitable for detection beyond $1.5 \mu\text{m}$. Recent work has included the development of intrinsic materials such as HgCdTe with narrower bandgaps. To date, however, there has been little application of these materials in IR astronomy. Extrinsic photoconductivity utilizes energy levels due to impurities that lie within the forbidden gap. Figure 2 illustrates the energy bands of a typical semiconductor

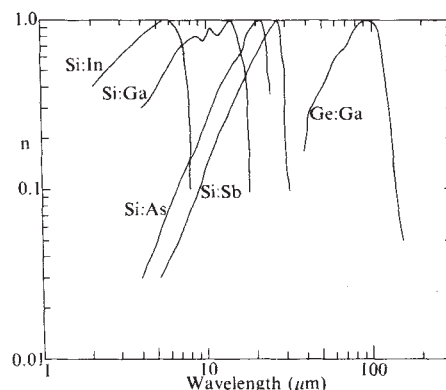


Fig.3 Normalized spectral response curves for some of the more important extrinsic detectors.

with a p-type dopant. In both silicon and germanium a wide variety of dopants with small ionization energies have been investigated⁸, and spectral responses of some of the more important ones are illustrated in Fig. 3. By choosing a suitable dopant, the spectral response of the detector can be tailored to the specific requirements of the experiment. Since extrinsic photoconductors have a definite long wavelength cut-off, they are not subject to noise due to thermal fluctuations in the cooling bath. As long as the temperature is low enough to prevent thermally excited carriers, the only important noise source is the background.

Extrinsic photoconductors are currently the most sensitive detectors available for operation under reduced backgrounds. Consequently, they are the principal detectors in the current generation of low background space IR astronomy experiments, including the Infrared Astronomy Satellite (IRAS) and the Spacelab 2 Infrared Telescope⁹ (SL-2 IRT). These space experiments will achieve greatly reduced background levels by cryogenically cooling the telescope. For example, the 12 μm band of IRAS will have $\sim 10^5$ times lower background power than with a comparable groundbased system. The better than 100-fold improvement in NEP translates to a 10^4 -fold reduction in integration time.

Current extrinsic photoconductors have typical quantum efficiencies of 30 per cent, although values as high as 80 per cent have been reported¹⁰. Noise equivalent powers in the 10^{-17} to 10^{-16} W Hz^{-1/2} range are attainable with TIA amplifiers and feedback resistors of $\sim 10^{10} \Omega$. These detector systems are adequate for the present generation of space experiments, where NEPs are limited by the thermal background from the interplanetary dust. Future IR astronomy experiments, such as the Shuttle Infrared Telescope Facility (SIRTF), possibly with diffraction-limited pixels and narrow bandwidths will require significant improvements in sensor technology to remain background-limited. In particular, amplification techniques will play a critical role in realizing the potential sensitivity.

A second exciting area of development is in large detector arrays. This technology has only recently become declassified, and much work remains to make arrays that are suitable for astronomical use. By the early 1990s, when SIRTF is hoped to be in operation, highly sensitive arrays of thousands of pixels should be available. □

X-ray astronomy

from Brian Taylor

THE very rapid advance of X-ray astronomy through the 1970s has produced demands for new detector developments with emphasis on large energy bandwidth (within the total range from tens of eV to tens of keV) spectral resolution, background rejection and position sensing capability if the momentum of the 1970s is to be maintained through the last two decades of the century. In particular position resolution is crucial given the advent of grazing-incidence imaging optics capable of providing arc second image quality at low energies, and coded aperture masks providing arc minute image quality at the upper end of the energy range.

Proportional counters

The state-of-the-art in mechanically collimated gas-filled proportional counters, the traditional work-horse for X-ray astronomy above 1.5 keV, is probably best exemplified by the eight, multiwire detectors of the Medium Energy Experiment¹ to be flown later this year on the EXOSAT mission². These detectors give an effective area of 1,600 cm², are collimated to 45 arc min full-width half maximum (FWHM) and cover the energy range up to 50 keV. Use of five-sided anticoincidence and risetime rejection techniques provide background rejection efficiencies >99 per cent in the range 1–6 keV, with an energy resolution of 22 per cent FWHM at 6 keV.

These collimated detectors, however, have a sensitivity limit of 0.3 millicrabs (10^{-11} erg cm⁻² s⁻¹) due to source confusion in the field of view. For the study of the spectral and temporal characteristics of the bright X-ray sources (from the UHURU or Ariel 5 catalogues) this is not a drawback, but for the determination of the high-energy spectral tails of weak sources detected by imaging telescopes better sensitivity is required. Conversely, the upper energy limit of current grazing-incidence optics (the Einstein observatory³, EXOSAT², ROSAT⁴) is typically a few keV and collimated proportional counters are unsuitable for the spectral mapping of extended objects (SNRs, clusters of galaxies) or resolution of confused regions.

This has led to the development of coded aperture mask telescopes for use with large area proportional counters with two-dimensional position sensing capability. Instruments to be flown⁵ on Spacelab 2 in 1984 and those considered for the potential ESA mission⁶ X-80 in 1988 are examples. In the case of Spacelab 2, the detector has a window area of 32 cm square, is filled with Xe-CH₄ to 1.5 atm

and with multilayer, multiwire arrays a position resolution of ≤ 1 mm is achieved, using delay line readout from orthogonal cathode wire planes. The anode wires provide the energy signal; an overall resolution of ~ 25 per cent being achieved at 6 keV. With a mask/detector separation of 3 m, mask elements 2.5 mm square and a field-of-view of about 5°, an angular resolution of ~ 3 arc min is achieved for a telescope covering the energy range from 2.5 to 25 keV — well suited to the study of temperature variations, iron line emission (near 7 keV) and non-thermal emission over clusters of galaxies.

Proportional counters with position sensing in essentially one dimension may be used in conjunction with Bragg Crystal Spectrometers. Such instruments were flown on the Solar Maximum Mission⁷ to study the solar X-ray spectrum, and Spherical Crystal Spectrometers⁸ are specified for the X-80 mission⁶ for the study of oxygen, silicon, sulphur and iron spectral features in cosmic X-ray sources. Resolutions $E/\Delta E$ of $\sim 10^3$ can be achieved, determined largely by crystal choice and geometry, and counters with a position resolution of ≤ 1 mm over a length of some 25 cm, achievable with the resistive wire anode, charge division technique are required. The spherical crystals allow the mapping of extended sources and position resolution in the plane orthogonal to the plane of dispersion is also required. The long, narrow counters require five-sided anticoincidence for good background rejection, given the relatively small crystal area contributing to a given energy and the long observation times of 10^4 – 10^5 s.

Proportional counters required to be sensitive above 1.5 keV normally employ thin ($\sim 50 \mu\text{m}$) Be foil entrance windows. For observations below 1.5 keV thin plastic windows must be used. Polypropylene has been widely used, for example the position sensitive proportional counter⁹ on EXOSAT has a 0.8 μm , 28 mm diameter window supported with a fine wire grid and withstands a pressure differential of 1.2 atm. However, to probe the XUV waveband at energies below 100 eV thinner windows are required and Lexan windows with thicknesses of $\leq 0.3 \mu\text{m}$ have been developed¹⁰.

At these low energies imaging telescopes are used to provide a large X-ray collecting area, requiring relatively small detectors. However, the planned Advanced X-ray Astrophysics Facility (AXAF)¹¹ with its 10-m focal length and 1.2-m diameter outer mirror requires an imaging proportional counter with a window aperture of 18 cm diameter to cover its 1° field of view.

Numerous methods have been developed for determining the position of the photoelectric absorption of X rays in

1. Herschel, W. *Phil. Trans. R. Soc.* **90**, 284 (1800).
2. Jones, R.C. *Proc. Inst. Radio Engrs.* **47**, 1459 (1959).
3. Mott, N.F. & Twose, W.D. *Adv. Phys.* **10**, 107 (1961).
4. Low, F.J. *J. opt. Soc. Am.* **51**, 1300 (1961).
5. Low, F.J. & Rieke, G.H. in *Methods of Experimental Physics: Astrophysics* (ed Carleton, N.) **12**, Part A, 415 (Academic, New York, 1974).
6. Hall, D.N.B. *et al. Appl. Opt.* **14**, 450 (1975).
7. Rieke, G.H. *et al. Appl. Opt.* **20**, 814 (1981).
8. Bratt, P.R. in *Semiconductors and Semimetals* (eds Willardson, R.K., Beer, A.C.) **12**, 39 (Academic, New York, 1977).
9. Young, E.T. *et al. Proc. S.P.I.E.* **280**, 12 (1981).
10. Young, E.T. & Low, F.J. *Proc. S.P.I.E.* **172**, 184 (1979).

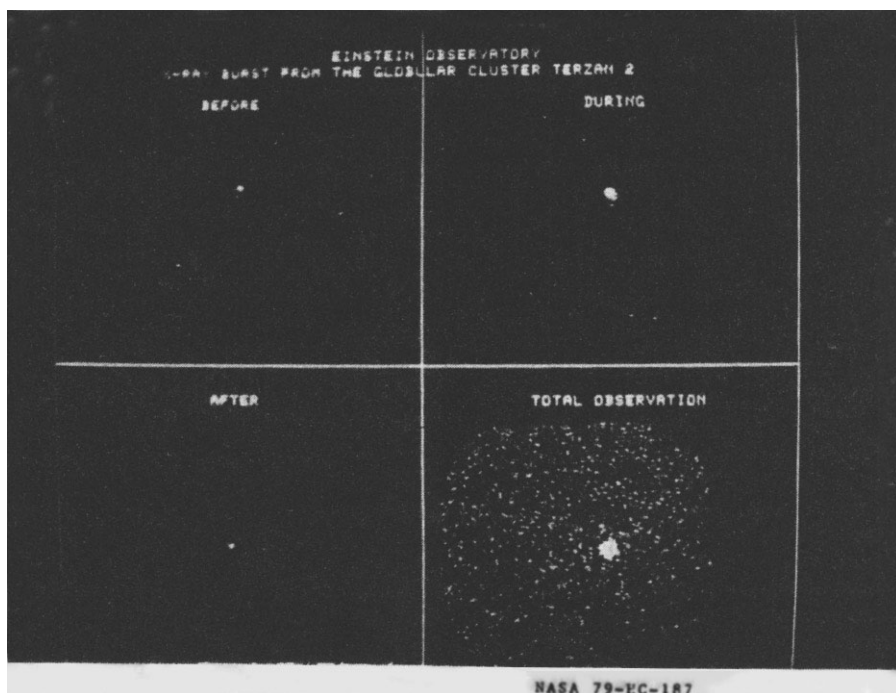
Brian Taylor is at the European Space Research and Technology Centre, Noordwijk, the Netherlands.

proportional counters in order to simplify the quantity of associated electronics. In the case of the Einstein IPC¹², the anode wire plane was sandwiched between two cathode planes, each comprising a resistive wire. A position resolution of 1.5 mm (FWHM) was achieved at 1.5 keV, corresponding to 1.5 arc min resolution through the optics. In the case of the EXOSAT PSD⁹, the anode is a resistive disk with four orthogonal pick-offs at the circumference. Pre-launch calibration of the argon/20 per cent methane filled detector has shown an on-axis position resolution of 0.2 mm (FWHM) at 1.5 keV, with a gas gain of 6×10^4 , equivalent to 40 arc s through the optics. The position resolution degrades off-axis with noticeable non-linearity, however, essentially distortion-free resistive anodes have been modelled¹³ and are under development. The energy resolution in the EXOSAT PSD is ~35 per cent FWHM at 1.5 keV.

Position accuracy in a proportional counter is ultimately limited by lateral diffusion of the primary electron cloud in the absorption region, but noise and the configuration of the readout system play significant roles, however a figure of 100 μ m FWHM at 1.1 keV has been achieved in a prototype XUV camera¹⁴. Such a detector performance coupled with a plate scale of the AXAF optics (50 μ m/arc s) would yield a spectral mapping capability down to 2 arc s.

Gas scintillation proportional counters

The energy resolution of practical proportional counters is typically 20–25 per cent at 6 keV with a $E^{-0.5}$ energy dependence. The statistics of the photoabsorption process and electron multiplication at the anode wire contribute about equally to the resolution. The past few years has seen rapid development of the gas scintillation proportional counter (GSPC) as a practical



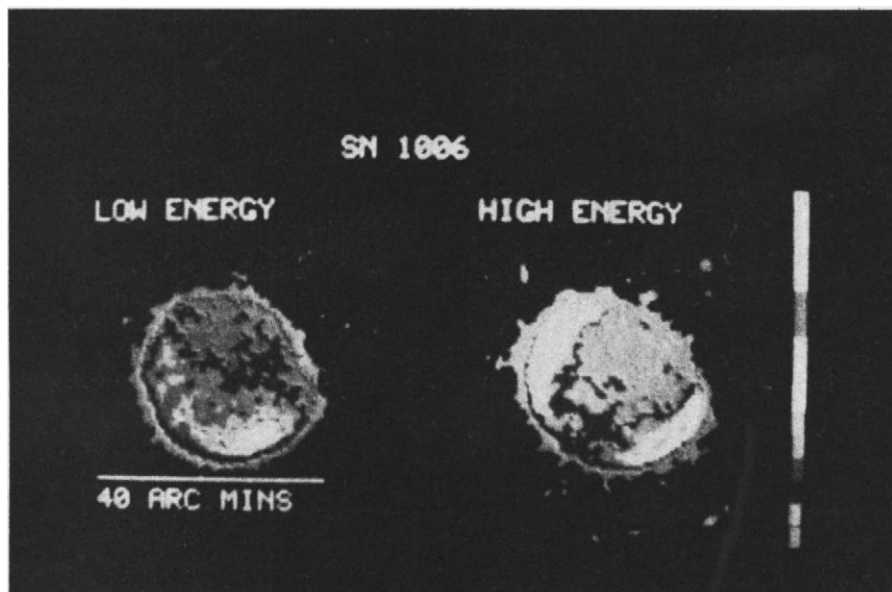
Sequence of 150-s observations of the X-ray burster in globular cluster Terzan 2 made with the Einstein Observatory IPC.

detector for spectroscopy in X-ray astronomy. In this device, electron multiplication is avoided; the primary photoelectrons are accelerated by an electric field, typically 4 kV cm⁻¹, through pure xenon. Copious quantities of UV photons are released from molecules, and are de-exciting following collision of the accelerating electrons with the gas. These UV photons (1,600–1,900 Å) can be observed directly with UV sensitive photomultipliers.

An energy resolution of 7.5 per cent FWHM at 5.9 keV has been achieved with a small non-imaging detector¹⁵ suitable for use with X-ray optics. The first flight of a 2 × 200 cm² effective area, spherical field

GSPC took place in September 1980 with a 180 second rocket observation of the SNR Cas-A¹⁶. These detectors with 10 per cent FWHM were able to resolve the iron line emission feature at ~7 keV from the continuum in the raw data — amply demonstrating their potential as broad-band spectrometers. Generically similar detectors will be flown on EXOSAT¹⁷, Spacelab 1 in 1983 (ref. 18) and the Japanese mission ASTRO-B¹⁹ (1983). Background rejection within these first generation detectors relies exclusively on UV burst length discrimination, analogous to risetime discrimination in conventional proportional counters. On the basis of laboratory simulations and the results of a preliminary rocket flight, it is estimated that the EXOSAT GSPC will achieve 97 per cent background rejection efficiency in the range 2–10 keV (ref. 17).

A number of methods have been examined to provide position sensing in GSPCs, including the use of multiple photomultipliers^{20,21}, on the principle of the Anger camera²², the coupling of a position sensitive photoionization detector (PID) using TMAE^{23,24} (tetakis [dimethyl-amine] ethylene with a photoionization potential of about 3.5 eV) and the coupling of a microchannel plate^{25,26}. In the two former cases, the energy resolution performance of the GSPC is maintained while position resolutions of about 1 mm at 6 keV have been achieved. The first rocket flight of a GSPC/PID coupled to X-ray optics took place in 1981 yielding multicoloured images of the Cygnus loop from which the distribution of oxygen and temperatures across the shock-front could be determined²⁷. Preliminary work with the coupled microchannel plate²⁶ and the use of coded TMAE as a liquid photocathode



X-ray image of the supernova remnant SN1006 (Einstein Observatory IPC), with selected energy range 0.01–0.3 keV (left) and 1.2–4.0 keV (right). (X-ray Astronomy Group, University of Leicester/Harvard-Smithsonian Center for Astrophysics.)

in conjunction with a very low pressure proportional counter²⁸, indicates that 300–500 μm position resolution at 6 keV should be achievable. This apparently modest position resolution, however, has to be seen in the context of the off-axis performance of the associated optics. Indeed, the half energy width of the AXAF optics could be as much as 1 arc min at the edge of the field of view¹¹.

The problem of background contamination becomes very severe for observations above 20 keV where up to now the 'Phoswich' detector has been employed largely in balloon-borne instrumentation. Such an instrument was successful in discovering the spectral feature in Her X-1 at ~60 keV ascribed to cyclotron emission²⁹. Alternatives to the phoswich are being developed including position sensitive proportional counters³⁰ and GSPCs³¹ for use with coded aperture masks. These detectors are filled up to 5 atm of xenon to provide adequate stopping power and incorporate 'fluorescent gating', in which an event qualifies only if the initial photo-absorption is detected in conjunction with the detectors of a xenon k_α or k_β fluorescent photon. With fluorescence gating applied to a non-imaging GSPC a background rejection of between 99.3 and 99.8 per cent was achieved over the energy range 35–100 keV coupled with an energy resolution of 2.7 per cent FWHM at 60 keV (ref. 32).

Recently the intriguing possibility of combining the spatial resolution of proportional counters with the energy resolution of GSPCs has been investigated³³. By employing a low drift field in a parallel plate proportional counter, the electron cloud is allowed to diffuse so that each individual electron produces its own independent avalanche, the light pulse from which can be detected and counted. As in the GSPC, energy resolution is deter-

mined by the statistics of the initial photo-absorption event and 19 per cent FWHM at 1.5 keV has been measured. Diffusion-limited position resolution has been achieved with an image intensifier.

Microchannel plates

The lower energy limit for gas-filled detectors is, of course, determined by the entrance window. Measurements into the EUV and XUV bands may be made using imaging optics in conjunction with the microchannel plate³⁴, an array of small (10–50 μm) channel electron multipliers. An energetic photon incident at the entrance to a channel liberates an electron which initiates an avalanche through secondary emission from a continuous dynode on the inside of a channel. By putting two or more plates in series and employing different bias angles (the angle of the channel to the plate surface) high charge gain (up to 10^8) can be achieved without ion feedback. Because the quantum efficiency of a microchannel plate is low it is customary to coat the surface of the front plate. In the case of the EXOSAT channel multiplier array (CMA) focal plane detector⁹, a magnesium fluoride coating is used which provides 30 per cent efficiency at 0.05 keV falling to 10 per cent at 1 keV, much smaller than the position sensitive proportional counter.

Several methods have been developed for MCPs including charge division readout of crossed wire grids as in the case of the Einstein observatory's high resolution imager³⁵, the resistive disc readout method for the EXOSAT CMA⁹ and various configurations of multi-anode readout arrays³⁶. The position resolutions achieved with MCPs are independent of energy and are largely governed by the readout method. It is anticipated that 15 μm will be achieved in the case of the AXAF HRI

(≈ 0.3 arc s) while 50 μm (≈ 10 arc s) is achieved by the EXOSAT CMA (plate diameter 50 mm, channel diameter 50 μm). Due to its saturated mode of operation, the microchannel plate has no intrinsic energy resolution, and offers the ultimate in 'black and white' imaging.

Focal plane detectors developed to date have been planar in section. However, the focal surface of grazing-incidence X-ray optics tends to be far from planar as one goes off-axis. The microchannel plate, by virtue of its construction, may be formed to adopt the contour of the focal surface. Such a development may be pursued for the wide-field, XUV camera of the ROSAT mission³⁷.

The detectors described above, the proportional counter, the gas scintillator and the microchannel plate are by no means the only ones employed or with potential in X-ray astronomy. Charge coupled devices (CCDs) have been adapted for the X-ray waveband³⁸, and with sizes of up to 25 mm square may achieve 15 μm position resolution and energy resolution comparable to the GSPC. A number of solid state detectors without position sensing properties and available in relatively small areas but with energy resolutions of 'a few per cent' cover the range from 1–100 keV, and include cooled Si(Li), intrinsic germanium³⁹, and mercuric iodide detectors⁴⁰, and finally there is the coded mask/sodium iodide crystal/Anger camera combination studied for the SIGMA mission⁴¹.

Operation of the proportional counter

To be detected in a proportional counter an X-ray photon must pass through the window and be absorbed in the gas inside the device. All the photon's energy is used in producing electron-ion pairs and the number of electrons liberated is proportional to the energy of the incident photon. The negatively charged electrons drift towards the anode wires under the influence of the electric field. Anode wires with radii of 10–30 μm are used so that the electric field near the wires is large enough to cause the electrons to ionize the gas molecules further. This process causes an avalanche of electrons, and the number of electrons reaching the wire may be several thousand times the number produced in the initial ionization. This multiplication factor is independent of the original number of electrons and so the size of the pulse sensed by electronics

connected to the anode wire is "proportional" to the energy of the incident photon. However, because of statistical variations in the initial ionization and avalanche, the response of a counter to X rays of a single energy is broadened, resulting in a finite energy resolution.

With a more complicated internal configuration and sophisticated electronics the location of the avalanche within the counter may be determined. By placing such a counter at the focus of an X-ray telescope it is therefore possible to obtain X-ray images. Unfortunately proportional counters are prone to many unwanted effects that may greatly degrade their performance. For this reason the choice of gas filling (usually based on an inert gas such as argon or xenon), and physical characteristics, are quite critical.

1. Turner, M.J.L. *et al. Space Sci. Rev.* **30**, 513 (1981).
2. Taylor, B.G. *et al. Space Sci. Rev.* **30**, 479 (1981).
3. Giacconi, R. *et al. Astrophys. J.* **230**, 540 (1979).
4. Aschenbach, B. *et al. Space Sci. Rev.* **30**, 569 (1981).
5. Willmore, A.P. *et al. Space Sci. Rev.* **30**, 601 (1981).
6. Taylor, B.G. *et al. Proc. COSPAR Symp.*, Ottawa (1982).
7. Rapley, C.G. *et al. Rev. Sci. Instrum.* **48**, 1123 (1977).
8. Schnopper, H. *et al. Space Sci. Rev.* **30**, 607 (1981).
9. de Korte, P. *et al. Space Sci. Rev.* **30**, 495 (1981).
10. Huizenga, H. *et al. Rev. Sci. Instrum.* **52**, 673 (1981).
11. Zombeck, M.V. *Proc. COSPAR Symp.*, Ottawa (1982).
12. Gorenstein, P. *et al. IEEE Trans. nucl. Sci.* **NS-28**, No.1, 869 (1981).
13. Frazer, G.W. *et al. Nucl. Instrum. Meth.* **190**, 53 (1981).
14. Huizenga, H. thesis, Univ. Amsterdam (1980).
15. Peacock, A. *et al. Nucl. Instrum. Meth.* **169**, 613 (1980).
16. Andresen, R.D. *et al. Space Sci. Rev.* **30**, 243 (1981).
17. Peacock, A. *et al. Space Sci. Rev.* **30**, 525 (1981).
18. Andresen, R.D. *et al. Proc. COSPAR Symp.*, Ottawa (1982).
19. Inoue, H. *et al. Paper presented at 1981 INS int. Symp. on Nuclear Radiation Detectors*, Tokyo (1981).
20. Hoan, N. *Nucl. Instrum. Meth.* **154**, 597 (1978).
21. Davelaar, J. *et al. Space Sci. Rev.* **30**, 591 (1981).
22. Anger, H.O. *Rev. Sci. Instrum.* **29**, 27 (1958).
23. Anderson, D.F. *IEEE Trans. nucl. Sci.* **NS-28**, **1**, 842 (1981).
24. Ku, W.H.-M. & Hailey, C.J. *IEEE Trans. nucl. Sci.* **NS-28**, **1**, 830 (1981).
25. Hailey, C.J. *et al. Nucl. Instrum. Meth.* **184**, 543 (1981).
26. Taylor, B.G. *et al. Proc. COSPAR Symp.* Ottawa (1982).
27. Ku, W.H.-M. *et al. Proc. COSPAR Symp.* Ottawa (1982).
28. Anderson, D.F. *Phys. Lett.* (submitted).
29. Trümper, J. *et al. Ann. N.Y. Acad. Sci.* **302**, 538 (1977).
30. Barbareschi, L. *et al. Proc. COSPAR Symp.* Ottawa (1982).
31. Taylor, B.G. *et al. IEEE Trans. nucl. Sci.* **NS-28**, **1**, 861 (1981).
32. Sims, M.R. *et al. Nucl. Instrum. Meth.* (submitted).
33. Siegmund, O.H. *et al. Proc. COSPAR Symp.*, Ottawa (1982).
34. Timothy, J.G. & Lapson, L. *Appl. Opt.* **13**, 1417 (1974).
35. Kellogg, E. *et al. Rev. Sci. Instrum.* **47**, 282 (1976).
36. Timothy, J.G. *et al. IEEE Trans. nucl. Sci.* **NS-28**, **1**, 689 (1981).
37. Pounds, K.A. *SERC Bull.* **2**, 14 (1982).
38. Dardas, D. *et al. Rev. Sci. Instrum.* **49**, 1273 (1978).
39. Proctor, R. *et al. Nucl. Instrum. Meth.* (in the press).
40. Vallerger, J. *et al. IEEE Trans. nucl. Sci.* **NS-29**, (1982).
41. Debouzy, G. *Proc. COSPAR Symp.*, Ottawa (1982).

REVIEW ARTICLE

The inflationary Universe—birth, death and transfiguration

John D. Barrow

Astronomy Centre, University of Sussex, Brighton BN1 9QH, UK

Michael S. Turner

Astronomy and Astrophysics Center, University of Chicago, Chicago, Illinois 60637, USA

A number of exciting new experimental and theoretical developments have recently made a dramatic impact on the interdisciplinary study of cosmology and elementary particle physics. The Nuffield Foundation recently sponsored a 3-week workshop in Cambridge where theoreticians worked intensively on some of these problems. A brief review is given of the progress that was made as a result of the workshop, something of the background to the problems tackled, and the likely course of future developments.

FOR the past four years the Nuffield Foundation has sponsored a workshop devoted to related aspects of gravitation and particle physics. Last month's workshop, organized by Stephen Hawking and Gary Gibbons at Cambridge, focused on the theory of the very early Universe (events occurring during the first second of its expansion history). The spectrum of participants was international with five particle physicists and cosmologists from the Soviet Union and the remaining 25 from Europe and the United States. It transpired that the conference organizers could not have chosen a better topic for detailed collective scrutiny. During the past six months there have been a number of dramatic theoretical and experimental developments on the interface between particle physics and cosmology that threaten to revolutionize our thinking about the large-scale structure of the Universe and link together our studies of the very large and the very small more closely than ever before. Early this year Linde¹, in the Soviet Union, Albrecht and Steinhardt² in the United States and Hawking and Moss³ in the UK proposed a model of the very early Universe (dubbed the 'new inflationary Universe') in which, through a phase transition soon after the big bang ($t \sim 10^{-35}$ s), a number of longstanding and outstanding cosmological problems seem to be simultaneously and naturally resolved. Then, last March, Cabrera of Stanford University, reported⁴ experimental evidence consistent with the detection of a single magnetic monopole. If, and this has not yet been confirmed, he has indeed seen the magnetic monopole predicted to exist by new grand unified gauge theories⁵ of elementary particle interactions at high energy (GUTs), then his result could provide a most dramatic and direct witness to the fact that the Universe was once at least as hot as 10^{27} K. Just to make things more interesting, it appears that Cabrera's monopole is inconsistent with the existence of the Galaxy's magnetic field; if these monopoles exist they should have discharged large-scale cosmic magnetic fields long ago⁶⁻⁸. In addition, monopoles would themselves provide microscopic 'laboratories' for the study of particle physics down at length scales $\sim 10^{-25}$ cm and a confirmation of the overall philosophy of grand unification. During the past year there has also emerged a growing interest in particle physics theories that not only unify the description of the three basic interactions (strong, weak and electromagnetic) but which are also supersymmetric (supersymmetry is a symmetry which relates the particles of integer spin (bosons) which do not obey the exclusion principle to those of half-integral spin (fermions) which do). The cosmological implications of such schemes are substantial but, so far, only partially explored. Almost all the

work done and talks presented at the workshop involved aspects of one or more of these three exciting developments.

Symmetry breaking and unification

The study of the very early history of the Universe amalgamates two seemingly unrelated disciplines: cosmology and elementary particles. So far, gauge theories have been very successful in describing and predicting the behaviour of fundamental particle interactions: the strong (or colour) force controlling quarks is successfully described by an unbroken SU(3) symmetric gauge theory; the weak and electromagnetic forces are conjoined in an SU(2) $\times$ U(1) symmetric gauge theory (the Weinberg-Salam-Glashow model) which is spontaneously broken at an energy of about 10^3 GeV. The point of spontaneous symmetry breaking (SSB) is that although the laws of physics may be intrinsically symmetrical, that symmetry is not manifest below a certain temperature or energy level because the lowest energy state of the system (the vacuum, as it is called) is a particular solution of the equations that does not possess their symmetry. In spontaneously broken gauge theories the properties of the vacuum state are of vital importance and are described by the vacuum expectation value of a scalar field—the Higgs field⁹. The appearance of a non-zero vacuum expectation value signals SSB.

The mathematical elegance and experimental vindication of the SU(3) and SU(2) $\times$ U(1) models has led to their incorporation within proposals for the grand unification of the strong and electro-weak interactions. In such models a spontaneous breakdown of complete symmetry between the strength and properties of these three interactions occurs when the temperature falls to 10^{28} K, an energy of about 10^{15} GeV. Clearly, if we are going to check such ideas by experiment we need energies of $\sim 10^{15}$ GeV to probe the symmetric vacuum. Such energies are never likely to be attained by terrestrial accelerators; so far only energies up to 1,000 GeV have been attained. There are very few low-energy consequences of GUTs that we can test but there are two generic predictions which are both extraordinary and accessible to current technology. It is predicted that the proton will be unstable with a lifetime $\sim 10^{31 \pm 2}$ yr and that stable, magnetic monopoles exist with a mass $\sim 10^{16}$ GeV (that is, about 10^{-8} g). At present about a dozen experiments are either operational or soon to be so and tuned to search for the characteristic proton decay events¹⁰. The Indian-Japanese collaboration, which is examining the stability of protons within about 140 tonnes of iron buried

below the Earth's surface in the Kolar Gold Fields, has already reported six candidate events for proton decay^{11,12} within the last two years, corresponding to a lifetime of about 10^{31} yr. Also, in the last month another claim to detect proton decay has been made by the CERN group experimenting inside Mt Blanc. They have detected a single candidate event.

The early Universe was not subject to the type of budgetary and technological limitations that we encounter today. According to the hot big bang model, temperatures corresponding to average particle energies as large as 10^{19} GeV should have been reached in the early Universe. So, it is clear that the early Universe provides an ideal 'theoretical laboratory' for the study of grand unification. By way of a corollary, grand unification can have remarkable consequences for the expansion and evolution of the Universe. Already, a key cosmological contribution has been made¹³⁻¹⁹: the same interactions that mediate proton decay today allow an initial universal state which is symmetric between matter and antimatter to evolve into one with an excess of matter which translates, after the antimatter annihilates, into one like our own that contains, on average, about 10^{10} photons for every nucleon of matter. Grand unified theories are able to explain both the origin of matter (rather than just radiation), the non-existence of antimatter in space, as well as the relative abundance of photons and nucleons—previously all unsolved puzzles.

Birth of the inflationary Universe

In 1981 Guth²⁰ proposed a new picture for the early stages of the hot big bang model which provided a 'natural' explanation for a collection of cosmological conundrums. The picture was christened the 'inflationary Universe'. Guth pointed out that if the Universe got 'caught' in the symmetric vacuum early on, when the temperature was below that at which the asymmetric vacuum was energetically more favourable (a lower energy state), then because of the large energy density associated with this state, the Universe would expand exponentially and 'super-cool', erasing its previous history. (Such an exponential expansion was first discussed by de Sitter²¹ for other, mathematical reasons.) When the transition to the asymmetric vacuum state does occur, an enormous 'latent' heat (the energy difference between the two vacuum states) is released, reheating the Universe so its subsequent evolution is that of the standard hot big bang model. As a result of this 'de Sitter' phase in its early evolution, the portion of the Universe that is observable today should be extremely uniform and expanding at a critical rate (that is, the ratio of the potential to the kinetic energy in the Universe, Ω , should be equal to 1). Without this de Sitter phase, in which the size of the Universe is greatly 'inflated' over what one would expect early on, the present uniformity of the Universe and the proximity of its expansion rate to the critical value remain a mystery. Our only 'explanation' would be to appeal to very special initial conditions which does not, of course, explain anything. Guth's inflationary Universe dispensed with the need for special initial conditions. Furthermore, it solved an embarrassing problem: the over-production of magnetic monopoles during the SSB phase transition of the GUT. In the earlier, more conventional models of the GUT phase transition so many monopoles were predicted to arise that they would contribute a density more than 10^{12} times larger than the maximum allowed by the observed deceleration of the Universe²²⁻²³. In the inflationary model, the period of accelerating expansion during the de Sitter phase dilutes the monopole density to a small and observationally permissible level. Unfortunately, as Guth and others realized, his original inflationary scenario possesses one fatal flaw: models in which the Universe gets trapped in the symmetric phase early on, end up with it caught there forever! If we are certain of nothing else, we can be sure that we do not live in a symmetric GUT vacuum—if we did then our protons would all decay in seconds.

Death and transfiguration

Six months ago Linde¹, Albrecht and Steinhardt² and Hawking and Moss³ analysed a different class of GUTs (those which undergo SSB of a characteristic type first studied by Coleman and E. Weinberg²⁴), and discovered that in these models the advantageous features of Guth's inflationary model could be retained whilst the difficulty of escaping from the symmetric vacuum of the de Sitter phase could be overcome. In their second generation model—the so-called 'new inflationary model'—the Universe never gets trapped in the symmetric vacuum state but instead simply takes a very long time to evolve from the symmetric to the asymmetric vacuum state, and while the Universe is evolving from the symmetric to the asymmetric state it expands exponentially due to the large energy density of the symmetric vacuum. In this new model the evolution to the true vacuum takes long enough for sufficient exponential expansion to occur to explain the uniformity, expansion rate and monopole-free composition of the present-day Universe.

Guth, Hawking, Linde, Moss and Steinhardt were all in attendance at the Nuffield Workshop and because of its potential explanatory power and topicality the 'new inflationary Universe' was the focal point of considerable critical scrutiny during the 3-week period. One exciting issue was actively investigated by four different research groups throughout the workshop (J. Bardeen, P. Steinhardt and M. Turner; A. Starobinskii; S. W. Hawking; A. Guth). They sought to calculate the form of the matter inhomogeneities that are spontaneously generated during the de Sitter phase and decide whether they explain the pattern of irregularity we see throughout the Universe today in the form of galaxies and clusters. In the new inflationary model the Universe emerges from the de Sitter phase too smooth to allow the subsequent development of objects like stars and galaxies. The four groups all considered the form of the density inhomogeneities that arise from spatial fluctuations in the scalar field that leads to the SSB of the GUT. Although by the second week of the workshop there was radical disagreement amongst the groups as to the magnitude of the final fluctuations—but complete agreement, both amongst themselves and with observation, as to the relative change in the magnitude from one length scale of the Universe to the next—by the end they were in agreement, and by very different methods had arrived at essentially the same result (testimony to the stimulating atmosphere of the programme). The result was bad news for the 'new inflationary Universe'—or at least for the simplest version studied so far: the resulting inhomogeneities were found to be far too big to lead to galaxies, rather everything would evolve to form superdense objects or black holes—not the type of Universe we live in. Linde²⁵ also reported that fluctuations in the scalar field might actually prevent the phase transition from occurring slowly enough to resolve the cosmological puzzles in the first place; again in the simplest model. The simplest model of new inflation was officially pronounced dead, just 6 months old.

However, the work done on this problem before and during the Cambridge meeting has been crucial. It has spelt out exactly what type of GUT is needed to have both a de Sitter phase and a resulting inhomogeneity spectrum in accord with present astronomical observations. By the end of the workshop Bardeen, Steinhardt and Turner had proposed a candidate supersymmetric model which satisfied these necessary criteria. At the moment, the basic idea of 'new inflation' still seems to possess attractive features. If an even 'newer' inflationary model can be constructed which is both a phenomenologically acceptable particle physics theory and without cosmological defect, then the marriage of GUTs and cosmology will have achieved a remarkable consummation. At a stroke a cluster of the most fundamental cosmological problems could be solved.

Lest anyone get carried away on the wave of euphoria surrounding inflation, the final summary talk of the workshop by Wilczek drew together several critical assessments that had been made earlier. The inflationary models 'predict' that today

the density parameter Ω should be equal to unity to a very high degree of precision; certainly no measurements we can now make should be able to measure the difference from unity. But, as Rees and Efstathiou reminded the particle theorists, the best astronomical determinations of Ω all consistently suggest a much smaller value, less than 0.2. A possible way out of this conflict is if the Universe is dominated ($\Omega \sim 1$) by non-baryonic matter which coalesces only on very large scales so that the observations made to date would not have been sensitive to the presence of this matter. Candidates for the closure density of dark matter include massive neutrinos, or one of the zoo of new particles postulated in supersymmetric gauge theories. Earlier in the week, Sciama⁵⁰ had stressed the attractiveness of photinos (supersymmetric fermionic partners of the photon) for providing the closure density and dark matter in the galaxy.

In the inflationary models the only relic monopoles which are present in the Universe today should be those produced shortly after the de Sitter phase by particle collisions^{26,51} and their abundance should be very small. This was one of the original virtues of the inflationary scenario. However, if Cabrera's observation of a monopole is correct and is representative of the typical flux in the Milky Way, then his observation appears to be in conflict with the inflationary model. Finally, Wilczek stressed that underlying the inflationary model is the value of the energy density associated with the vacuum. At present, particle physics calculations do not set the absolute scale of this energy density, only the difference in energy between the symmetric and asymmetric vacua. Our observations of the total mass density of the Universe show that, compared to this energy difference, the energy density of our asymmetric vacuum must be at least a factor 0 (10^{20}) smaller. If we define the energy density of the symmetric vacuum as the zero level then we can calculate the energy density of the asymmetric vacuum. Unfortunately, at present, we have no idea why the energy density of our vacuum should turn out to be so infinitesimal. An understanding of this 'fact' or a recasting of the problem ought to reveal a deep connection between particle physics and gravity—and that may even spring a few surprises of its own.

Monopoles

Ever since 1931, when Dirac²⁷ showed that the existence of magnetic monopoles requires the quantization of both electric and magnetic charge, physicists have continually searched for (and occasionally even claimed to have found) magnetic monopoles. The advent of GUTs has revived interest in monopole-hunting. These theories always predict the existence of magnetic monopoles with a mass close to the energy scale of grand unification^{28,29}. Last March, Cabrera⁴ reported experimental evidence consistent with the detection of a single monopole. He utilizes a SQUID to monitor the current in a superconducting loop surrounded by a superconducting magnetic shield which in turn enables him to detect magnetic charges. If a magnetic charge passes through the loop it induces a persistent current whose amplitude is proportional to the magnetic charge. Unfortunately, sources of noise and systematic error might also induce such flux jumps and only time will tell whether this event is a real monopole. Cabrera's event has the Dirac charge value and corresponds to a 'flux' (one event in 150 days) of about $10^{-9} \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$. In his review of the 'monopole problem' at the workshop Preskill described several new types of experiment planned by Cabrera and others which will both increase the effective capture area of the detector and use independent cross-checks to eliminate many possible sources of systematic error. Within a year his 'observed flux' should be confirmed or disproved. Unfortunately, Cabrera's experiment (as well as all the other monopole searches planned or in operation) can give no indication of the monopole mass. To measure that, its path through a very strong magnetic field would have to be tracked very accurately as superheavy monopoles are not easily deflected from straight lines.

As mentioned above, if this measurement is really of a magnetic monopole, its implications for cosmology and particle physics are extremely significant. Astrophysically, Cabrera's flux contradicts everything we know about the magnetic field of the Galaxy. Monopoles should move along magnetic field lines in the Galaxy, thereby gaining kinetic energy at the expense of the magnetic field energy. A monopole flux in the Galaxy exceeding $\sim 10^{-15} \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ (the so-called Parker⁶⁻⁸ limit) would rapidly dissipate the entire galactic field on a time scale shorter than that on which it can be regenerated. Our observation of an interstellar field of a few microgauss means that Cabrera's flux cannot be indicative of the average galactic flux of monopoles. However, we clearly occupy a special position in the Galaxy³⁰—very close to a star, the Sun. Wilczek and Preskill discussed the possibility that the average galactic flux is close to, or less than $\sim 10^{-15} \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ but the Sun captures a cloud of monopoles by gravitational attraction leading to a flux in the solar neighbourhood that exceeds the average value in the Galaxy by a factor that can be as large as the required value $\sim 10^6$.

The GUT monopole is a remarkably interesting object. It has a size close to that of other elementary particles but an enormous mass (by microscopic standards), close to that of a bacterium. This means monopoles are not easily stopped and they pass through the Earth without significant energy loss. The monopole interior and outlying periphery has an onion-skin structure. Very close to its centre the vacuum is the GUT symmetric vacuum, a little farther out it is the $\text{SU}(2) \times \text{U}(1)$ symmetric vacuum. A single monopole thus provides a potential laboratory for studying grand unification and SSB.

The rich internal structure of a monopole also means that physics close by is likely to be extremely interesting. Callan³¹, Rubakov³² and Wilczek³³ have all pointed out that monopoles should catalyse proton decay, for example, via $M + p \rightarrow M + e^+ + X$ where X = other particle debris. Callan and Rubakov argued that the cross-section for these exchanges should be large, comparable to a typical hadronic cross-section ($\sim m_\pi^{-2}$). At the workshop Ellis³⁴ discussed an interesting terrestrial consequence of this: terrestrial protons exposed to a Cabrera monopole flux should, on the average, decay with a lifetime of $\sim 10^{28} \text{ yr}$. The present experimental evidence on proton places a limit on the proton lifetime, $\geq 3 \times 10^{30} \text{ yr}$, which rules this out. Preskill and Wilczek also pointed out that if the Callan-Rubakov estimate is correct then neutron stars would present an even more serious problem^{35,36}. Neutron stars are very effective at capturing monopoles because of their high densities and gravitational field strengths. Once captured by the neutron star, the monopoles would begin catalysing neutron decay and, unless the monopole flux were less than 10^{-8} of Parker's bound (and 10^{16} times smaller than Cabrera's flux) neutron stars would emit far more X-ray and UV radiation than is observed. This leads to an unfortunate conclusion: if the cross-section for this interesting process is large, then the flux of monopoles must be undetectably small.

However, Preskill and Wilczek sharply disagree with the Callan-Rubakov estimates of the cross-section for monopole catalysis of proton decay. They presented convincing arguments that the cross-section should be a weak interaction cross-section, or something even smaller, and this means it is at least 10^{10} smaller than the estimates of Callan and Rubakov. The problem of computing this catalysis cross-section is a formidable mathematical task and its value must still be regarded as uncertain.

Monopoles are not the only esoteric structures that can emerge from the peculiar vacuum physics of GUTs. At the workshop Shafi and Vilenkin discussed the possibility that strings or walls of vacuum energy could be produced at phase transitions where SSB occurs and these structures could persist until the present³⁷⁻⁴⁰. (Strings, walls and monopoles are all topological structures associated with the direction in group space that the Higgs field responsible for SSB has in different regions of our physical space.) Unlike in the case of monopole

production at the moment of SSB in GUTs there seems to be no problem in creating too many strings and walls. The possibility of wall structures appears to be ruled out by observation—the presence of one stretching across the observable Universe would create an anisotropy in the temperature distribution of the microwave background radiation much larger than is observed. Vilenkin argued that, by way of contrast, vacuum strings could have a very significant astrophysical role. A network of string-like structures might provide a natural explanation for the level of inhomogeneity necessary to create galaxies later on⁵². Also, the oscillation of strings that close upon themselves to form isolated loops could give rise to an observable flux of gravitational radiation. However, this conclusion seems to rest on an axiom of faith that the strings would dissipate their energy into gravitational waves rather than photons which would leave no trace if they were produced early enough in the history of the Universe. Finally, to give some feel for what a string is like, a vacuum string arising from the era of GUT would have a thickness of the order of the Compton wavelength of the Higgs scalars involved in the SSB ($\sim 10^{-27}$ cm) and a calculable mass per unit length such that a piece of vacuum string stretching 15 thousand million light years across the observable Universe would have a mass roughly equal to that of 10 million galaxies.

Supersymmetry

One of the most important unsolved problems of modern particle physics is the gauge hierarchy problem—the difficulty of constructing a theory that explains why SSB takes place on such widely separated energy scales, 10^3 GeV and 10^{15} GeV. One possible resolution which has been the focus of growing attention^{41–44} is super-symmetric grand unification (SUSY). Supersymmetry requires the existence of a bosonic partner of identical mass for every known fermion and vice versa. As no such supersymmetric partners of the known quarks and leptons exist (they would have been seen), supersymmetry, like some of the gauge symmetries, must be a broken symmetry. Supersymmetry has its own intrinsic appeal for other reasons²⁶: it is the only way to unify the description of particles with different spin statistics, and it is the only way to unify internal symmetries (that is, those associated with the gauge symmetries of particle physics) with the space-time symmetries of general relativity (and hence is a very good candidate for unifying gravity with the strong and electro-weak interactions). However, there appear to be a number of substantial cosmological hurdles in the path of supersymmetric GUTs. Ellis, Lazarides and Nanopoulos each discussed some of these difficulties and possible ways of overcoming them. In particular it seems to be difficult to arrange things so that our asymmetric vacuum is the lowest vacuum energy state, and even if this can be achieved, it is hard to guarantee that the Universe evolves into the asymmetric vacuum rather than the cul-de-sac of the symmetric state where it remains stuck forever. Furthermore, it is difficult to construct a grand unified SUSY model which generates the observed matter–antimatter asymmetry and specific entropy of the Universe because it is harder to get baryon non-conserving interactions out of equilibrium than in conventional GUTs. At present it is premature to accept or reject such models and SUSY remains an attractive idea for resolving the hierarchy problem and a possible pathway to a theory of quantum gravity since gravity can be incorporated when the supersymmetry is made a local rather than a global (as in SUSY GUTs) symmetry.

Gravitation

General relativity is a gauge theory although not a quantum theory but it has long been realized that quantum corrections to the classical equations of Einstein must arise at energies exceeding about 10^{19} GeV, or, in the cosmological problem at times, $t \leq 10^{-43}$ s. These energies exceed even those required for grand unification and as yet there has appeared no way of testing theories of quantum gravity by experiment. A full-blooded quantum theory of the gravitational field, including therefore the space–time geometry, is not yet on offer. A more modest programme is being undertaken to determine how quantum fields behave in a classical gravitational field^{45,46} and progress in this area was reviewed in Cambridge by Hartle and Hu. As yet there is no ‘standard’ model in this work. To illustrate this, consider the problem of how quantum effects determine the evolution of the Universe before 10^{-43} s. In some models the singularity of infinite density at zero time remains, whilst in other, equally likely variants, it is removed and the Universe bounces; the pre-bounce state also admits a variety of possibilities and it is still premature to try and draw definite conclusions as to the first instants of the big bang.

Two topics of interest to quantum gravity pundits were presented at the workshop: the first, by Starobinski⁴⁷, who showed how an inflationary phase of de Sitter expansion may have occurred at the Planck time, 10^{-43} s, rather than later at the time of grand unification. Of particular interest is the possibility that such an event might leave an observable imprint on the gravitational wave background in the Universe which we could one day detect. The other contribution by Novikov described how, close to the Planck epoch, the propagation of sound waves in the Universe may be described by a wave equation that is not conformally invariant. These sound waves can be amplified by interacting with the curvature potential created by the expansion of the Universe and give rise to a characteristic spectrum of irregularities today^{48,49}.

The fundamental cosmological puzzles—the expansion rate of the Universe, its isotropy, homogeneity, irregularity spectrum and specific entropy—all involve the very special initial conditions we seem to require at 10^{-43} s. It may be that these puzzles will be explained by events occurring at the Planck epoch rather than through inflation at the time of grand unification, and that a resolution of our quandaries must await the much sought synthesis of relativity and quantum field theories. What we have investigated so far is by no means the end but hopefully it is the beginning of a true story with a happy ending.

The consensus among the participants was that they had attended a significant and productive scientific gathering in which real scientific progress was made during its extent. The organizers and sponsors were largely responsible for this; without their efforts it would not have been possible for all the protagonists in the early Universe debate to meet and exchange ideas. The simplest ‘new inflationary’ Universe appears to be defunct; however, perhaps reports of its death ‘have been greatly exaggerated’ and a newer version will emerge to fulfil the criteria for success that were drawn up at the workshop. Important experimental and theoretical issues regarding the existence and behaviour of magnetic monopoles have emerged. Supersymmetric grand unification has been subject to significant scrutiny and has been pushed near the brink of viability. Many of these exciting developments have emerged only during the last year and no one has any doubt that this trend is going to continue for some considerable time.

We thank all the participants in the Nuffield Workshop for conversations and the Nuffield Foundation for support.

1. Linde, A. D. *Phys. Lett.* **108B**, 389–395 (1982).
2. Albrecht, A. & Steinhardt, P. J. *Phys. Rev. Lett.* **48**, 1220–1224 (1982).
3. Hawking, S. W. & Moss, I. G. *Phys. Lett.* **110B**, 35–38 (1982).
4. Cabrera, B. *Phys. Rev. Lett.* **48**, 1378–1382 (1982).
5. Langacker, P. *Phys. Rep.* **72C**, 185–385 (1981).
6. Parker, E. N. *Astrophys. J.* **160**, 383–404 (1970).
7. Bludman, S. A. & Ruderman, M. A. *Phys. Rev. Lett.* **36**, 840–843 (1976).

8. Lazarides, G., Shafi, Q. & Walsh, T. F. *Phys. Lett.* **100B**, 21–24 (1981).
9. Abers, E. S. & Lee, B. *Phys. Rep.* **9C**, 1–141 (1973).
10. Weinberg, S. *Scient. Am.* **244**(6), 64–75 (1981).
11. Krishnaswamy, M. R. *et al. Phys. Lett.* **106B**, 339–346 (1981).
12. Goldhaber, M. & Sulak, L. R. *Comm. Nucl. Part. Phys.* **10**, 215–227 (1981).
13. Sakharov, A. D. *Soviet Phys. JETP Lett.* **5**, 24–27 (1967).
14. Yoshimura, M. *Phys. Rev. Lett.* **41**, 281–284 (1978).

15. Ignatiev, A. Y., Krasinokov, N., Kuzmin, V. & Tavkhelidze, A. *Phys. Lett.* **76B**, 436–440 (1978).
16. Dimopoulos, S. & Susskind, L. *Phys. Rev. D* **18**, 4500–4509 (1978).
17. Toussaint, D., Treiman, S. B., Wilczek, F. & Zee, A. *Phys. Rev. D* **19**, 1036–1045 (1979).
18. Weinberg, S. *Phys. Rev. Lett.* **42**, 850 (1979).
19. Ellis, J., Gaillard, M. K. & Nanopoulos, D. V. *Phys. Lett.* **80B**, 360–364 (1978).
20. Guth, A. H. *Phys. Rev. D* **23**, 347–356 (1981).
21. de Sitter, W. *Mon. Not. Roy. astr. Soc.* **78**, 3–28 (1917).
22. Preskill, J. P. *Phys. Rev. Lett.* **43**, 1365–1368 (1979).
23. Zeldovich, Ya. B. & Khlopov, M. Yu. *Phys. Lett.* **79B**, 239–241 (1979).
24. Coleman, S. & Weinberg, E. *Phys. Rev. D* **7**, 1888–1910 (1973).
25. Linde, A. D. *Lebedev Phys. Inst. Preprint* (1982).
26. Turner, M. S. *Phys. Lett.* (in the press).
27. Dirac, P. A. M. *Proc. R. Soc. A* **133**, 160–172 (1931).
28. Polyakov, A. *Soviet Phys. JETP Lett.* **20**, 194–195 (1974).
29. 't Hooft, G. *Nucl. Phys.* **79B**, 276–284 (1974).
30. Dimopoulos, S., Glashow, S. L., Purcell, E. M. & Wilczek, F. *Nature* **298**, 824–825 (1982).
31. Callan, C. G. *Princeton Univ. Preprint* (1982).
32. Rubakov, V. S. *Soviet Phys. JETP Lett.* **33**, 644–646 (1981); *USSR Acad. Sci. Inst. Nucl. Res. Preprints* P-0204, P-0211 (1981).
33. Wilczek, F. *Phys. Rev. Lett.* **48**, 1146–1150 (1982).
34. Ellis, J., Nanopoulos, D. V. & Olive, K. A. *CERN Preprint* TH-3323 (1982).
35. Kolb, E. W., Colgate, S. A. & Harvey, J. A. *Phys. Rev. Lett.* (in the press).
36. Dimopoulos, S., Preskill, J. P. & Wilczek, F. (in preparation).
37. Kibble, T. W. B. *J. Phys.* **9A**, 1387–1398 (1976).
38. Zeldovich, Ya. B. *Mon. Not. R. astr. Soc.* **192**, 663–667 (1980).
39. Vilenkin, A. *Phys. Rev. D* **24**, 2082–2089 (1981).
40. Kibble, T. W. B., Lazarides, G. & Shafi, Q. *Phys. Lett.* **113B**, 237–239 (1982).
41. Dimopoulos, S. & Raby, S. *Nucl. Phys.* **192B**, 353–368 (1981).
42. Witten, E. *Nucl. Phys.* **188B**, 513–554 (1981).
43. Dine, M., Fishler, W. & Srednicki, M. *Nucl. Phys.* **189B**, 575–593 (1981).
44. Fayet, P. & Ferrara, S. *Phys. Rep.* **32C**, 249–334 (1977).
45. Birrell, N. & Davies, P. C. W. *Quantum Fields in Curved Space* (Cambridge University Press, 1982).
46. Isham, C., Penrose, R. & Sciama, D. W. *Quantum Gravity 2* (Oxford University Press, 1981).
47. Starobinski, A. A. *Phys. Lett.* **91B**, 99–102 (1980).
48. Kompanyets, D. A. & Lukash, V. N. *Soviet Astr.* **25**, 272–274 (1981).
49. Mukhanov, V. F. & Chibisov, G. V. *Soviet Phys. JETP Lett.* **33**, 532–535 (1981).
50. Sciama, D. W. *Phys. Lett.* **114B**, 19–22 (1982).
51. Preskill, J. P. *Harvard University Preprint* (1982).
52. Barrow, J. D. & Silk, J. *Astrophys. J.* **250**, 432–449 (1981).

ARTICLES

Phase transitions and dynamics of the Universe

Vahé Petrosian

Institute for Plasma Research and Department of Applied Physics, Stanford University, Stanford, California 94305, USA

The restoration of symmetry at grand unification in a closed contracting Robertson–Walker universe could slow down and halt the contraction, causing the universe to bounce and avoid the singular state or the big crunch. During subsequent expansion the symmetry is broken and the universe reverts back to its present (normal) state. A scenario is presented whereby the present entropy of the substance (resulting in $\sim 10^{90}$ photons) is created through a series of such bounces along with three possible initial states. The monopole, horizon and flatness difficulties faced in ordinary big bang models are also solved by the model presented.

DESPITE astronomical discoveries during the past half century, the basic and simple nineteenth century questions, such as age and size of the Universe, remain essentially unanswered. The open and flat models of the Robertson–Walker metric are unsatisfactory because of the need to extrapolate the local observations to infinity. The closed and finite (and therefore more appealing) model suffers from the fact that it will eventually collapse into a singular state. Although introduction of a positive and constant cosmological term will solve the collapse problem, there still remains (for all big bang models) the difficulty of dealing with the singular initial state which took place in a finite past. There is always the hope that quantum gravity will solve this problem around the Planck era (energies $\sim 10^{19}$ GeV). In this article, recent developments in microphysics are used to propose a possible scenario for the bounce of a collapsing closed universe occurring at the grand unification era (energies $\sim 10^{15}$ GeV) before it reaches the Planck era.

One of the intriguing results from the unification at high energies of various forces of nature is that as the expanding universe (or any system in thermal equilibrium) cools below some critical temperature the symmetry is spontaneously broken, the universe undergoes a first-order phase transition which induces a change in the vacuum energy density¹. Without considering the details of the microphysics, most of which is still open to speculation, I take a phenomenological approach and investigate the modification of the dynamics of a Robertson–Walker universe due to a transition from a higher (or false) to lower (or true) vacuum state and vice versa.

Outline of the problem

It is well known^{2,3} that a non-zero vacuum energy density E_v corresponds to a non-zero value of the cosmological constant $\Lambda = 8\pi G E_v / c^2$, whose effects on the dynamics of the universe

can be taken into account by redefinition of energy density and pressure as $E = E_s + E_v$, $P = P_s + P_v = P_s - E_v$, where E_s and P_s are the energy density and pressure of substance (matter and radiation^{4,5}). As proposed by Einstein, the cosmological term, or E_v , cannot be diverged. Variation in this term can be generated by a scalar field⁶. The possibility of a phase transition allows variation in the value of Λ or E_v . Then the conservation equations must be generalized to allow for the conversion of vacuum energy density into energy density of substance and vice versa⁷: $a(dE_s + dE_v) = -3(E_s + P_s)da$. Here a is the expansion parameter. We are interested in the early phase of the universe where E_s is dominated by radiation and relativistic pairs in thermal equilibrium. Then $P_s = E_s/3$ and $E_s = (1 + g)E_r$, where g (which can vary across a phase transition) is the contribution of fermion and boson pairs, and $E_r \propto T^4$ is the energy density of radiation at temperature T .

In what follows, modified Planck units $8\pi G/3 = \hbar = c = 1$ are used. Normally $S^{4/3}E_s$ is a constant and in thermal equilibrium, $S = (1 + g)^{3/4} 0.15 N_r$ (N_r is the total number of photons in the closed universe of volume $2\pi^2 a^3$), which is proportional to the total number of free particles in the universe, is also a constant or varies due to the changes in g . During a phase transition, however, because of presence of the dE_v term in the above conservation equation, this is no longer true and the differential equations governing the expansion parameter can be written as (overdot stands for derivative with respect to time t)

$$\begin{aligned} (\dot{a}a)^2 &= E_v a^4 - a^2 + S^{4/3} \\ \ddot{a}a^3 &= E_v a^4 - S^{4/3} \end{aligned} \quad (1a)$$

from which (or from the conservation equations directly) we obtain a modified form for the conservation laws.

$$dS^{4/3}/dt = -a^4 \dot{E}_v \quad (1b)$$

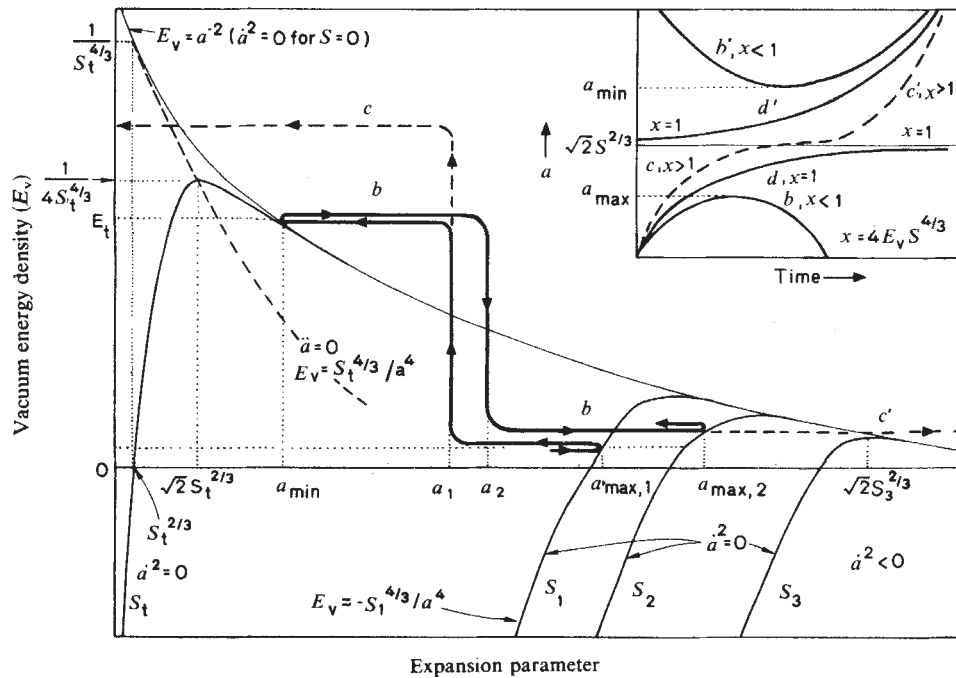


Fig. 1 Schematic variation of vacuum energy density with the expansion parameter for a radiation-dominated universe. Above the thin dashed line marked $\ddot{a}=0$ expansion is accelerated. Thin solid lines represent $\dot{a}^2=0$. For higher values of S these curves shift to the right and decrease in amplitude. For each S the region below this line ($\dot{a}^2<0$) is forbidden. The heavy solid and dashed lines show various paths of evolution with phase transitions at a_1 during the contraction and at a_2 during the expansion. The letters b, b', c, c' refer to the type of models described in the inset, where $x=1$ corresponds to the Einstein static model.

Conventional models

Let us first review briefly the relevant aspects of the usual closed models⁸ with $E_v=0$. Figure 1 shows curves of E_v against a for $\ddot{a}=0$ (thin dashed lines) and $\dot{a}=0$ (thin solid lines) at various values of S . The schematic variation of the expansion parameter a with time is shown in the inset of Fig. 1 for $x=4S^{4/3}E_v>1$ (models c, c'), $0<x<1$ (models b, b') and $x<1$ (models d, d' and the Einstein static model). Of particular interest are the models with $0<x<1$ where the universe has either a maximum or a minimum size:

$$\begin{aligned} a_{\max} &= (2E_v)^{-1} [1 - (1-x)^{1/2}] \\ a_{\min} &= (2E_v)^{-1} [1 + (1-x)^{1/2}] \end{aligned} \quad (2)$$

For a constant E_v (therefore a constant S) the expansion or contraction of the universe follows a horizontal path until it reaches an $\dot{a}=0$ curve (for the appropriate value of S) after which it reverses its motion contracting to a singularity or expanding to infinity (see models b and b' in the inset to Fig. 1). For large values of E_v ($x>1$) $\dot{a}$ never becomes zero and expansion continues forever.

Models with variable cosmological terms

If $\dot{E}_v \neq 0$, then the evolution of the universe will not be constrained along a horizontal line and at different stages will be limited by the $\dot{a}=0$ curves of different values of S . It is therefore clear that arbitrary variation of E_v (and the corresponding variation in S) could result in a complicated evolution of the universe. Here we examine the simple variation of E_v from a small value to a large value comparable with the energy density of substance at the time of the variation.

Let us assume that the universe is closed, $S=S_1$, and that the vacuum energy density $E_v=E_1<(4S_1)^{-4/3}$ ($\sim 10^{-120}$ for our present universe^{8,9}). The universe then will evolve as in model b (see inset to Fig. 1), expanding to a maximum size, $a_{\max}$, after which it begins to collapse, as shown by the heavy line directed to the left in Fig. 1. (Our present universe is matter dominated; therefore, it will reach a maximum value if $E_1<M^{-2}\sim 10^{-122}$ where $M=2\pi^2 a^3 \rho$ is the total mass of the non-relativistic particles in the universe. However, when it collapses to a size smaller than its present size by a factor of $\sim 10^3$, it will become a radiation dominated universe and revert back to our idealized model. Note that the present observational limits on E_0 do not rule out ever-expanding models such as model c' .) As the universe collapses it will follow an $E_v=\text{constant}$ track of a radiation

dominated universe and the temperature will rise as

$$T = 1.11/(1+g)^{1/4} S^{1/3}/a \quad (3)$$

until it reaches the critical temperature T_0 of the phase transition above which the vacuum energy density will increase by $E_0 \sim T_0^4$ ($\sim 10^{-20}$ for grand unification phase transition = GUT). After the phase transition is completed the vacuum energy density is $E_v=E_1+E_0$ and, according to equation (1b), the number of photons and pairs changes to S_1 given by

$$S_1^{4/3} - S_1^{4/3} = - \int_{E_1}^{E_1+E_0} a^4 dE_v \approx -a_1^4 E_0 \quad (4)$$

The last equality will be valid if the transition occurs at a_1 instantaneously, or on a time scale short compared with the Hubble time at that epoch which is about $E_v^{-1/2}$ ($\sim 10^{-10}$ for GUT). Otherwise, a_1^4 can be taken to be an average value of a^4 during the transition period. The fate of the universe after the transition depends on the value of S_1 and E_1 . Note that right after the transition the universe is still contracting

$$\dot{a} = -(E_1 a^2 - 1 + S_1^{4/3}/a^2)^{1/2}, \quad \ddot{a} = a E_1 - S_1^{4/3}/a^3 > 0 \quad (5)$$

Initially the first terms in these equations are dominant and the contraction will be exponential; $a \propto \exp(-E_1 t)$. However, as a decreases the other terms (due to curvature and substance) in these expressions become more important. Whether the universe will bounce or continue to contract will depend on which of these two terms in the expression for $\ddot{a}$ becomes important first. If S_1 is large, so that $x_1=4S_1^{4/3}E_1>1$, then the radiation term ($S_1^{4/3}/a^4$) will become important before the curvature term ($1/a^2$) and the universe will contract to a singularity as shown by the heavy dashed line (c) in Fig. 1. On the other hand, if S_1 is small so that $x_1<1$, then the curvature term becomes dominant first, and the universe will reach the minimum value of the expansion parameter given in equation (2) and then will expand with acceleration as in model b' of the inset and as shown by the heavy solid line directed to the right in Fig. 1. During this adiabatic contraction and expansion phase S_1 remains constant and the energy density of the substance increases but never exceeds E_1 . The evolution during the expansion phase may follow an inflationary scenario similar to Guth's and others¹⁰⁻¹² where the symmetry breaking is delayed, the universe supercools, but finally it undergoes the phase transition releasing the vacuum energy density which heats the universe back up to a temperature of the order of T_0 . However, in the model proposed here the expansion does not begin with

a high entropy state but instead is dominated by the curvature and vacuum energy density similar to the scenario proposed by Bludmann¹³.

Suppose this transition occurs at $a = a_2$, then the final value of S is given by

$$S_2^{4/3} = S_1^{4/3} + a_2^4 E_0 = S_1^{4/3} + (a_2^4 - a_1^4) E_0 \quad (6)$$

After the phase transition is completed the universe evolves as a radiation-dominated universe ($E_v = E_1 \ll E_s$, see model *b* in Fig. 1) and if $4E_1 S_2^{4/3} < 1$ it reaches a maximum expansion with $a_{\max,2} = S_2^{2/3}$. This cycle is then repeated if $4E_1 S_2^{4/3} < 1$ and $4E_1 S_1^{4/3} < 1$. These results are summarized in Fig. 2 where we plot schematic variations of a , E_v , $S^{4/3}$ and the three terms in the expression for $(\dot{a}/a)^2$ versus time for a bounce.

Note that there may be more than one spontaneous symmetry breaking and corresponding numbers of phase transitions, so that S_1 and S_2 are not necessarily the final numbers of photons in the universe. Furthermore, the bounce would occur only for grand unification (or for a higher temperature) phase transition when the net baryon to photon number ratio $N_B/N_\gamma \ll 1$. Otherwise, between t_1 and t_2 the universe never becomes curvature dominated and the net baryon term, N_B/a^3 (which, of course, cannot be eliminated during any phase transitions occurring at temperatures less than the grand unification temperature), remains more important than the curvature term, $1/a^2$. If N_B is not zero, then it can be shown that a bounce becomes possible if both $S_i < E_0^{-3/4}$ ($\sim 10^{15}$ for GUT) and $N_B < E_0^{-1/2}/m_B$ ($\sim 10^9$ for GUT). Here $m_B \sim 10^{-19}$ is the rest mass of the baryons. Similarly, our scenario requires that near the bounce the curvature term must also dominate terms due to anisotropies¹⁴ or other sources.

Finally, if the process of symmetry restoration and breaking is not reversible, the two transitions may occur at different values of the expansion parameter. In particular, if $a_2 > a_1$, then $S_2 > S_1$ and entropy is generated during a bounce. This suggests

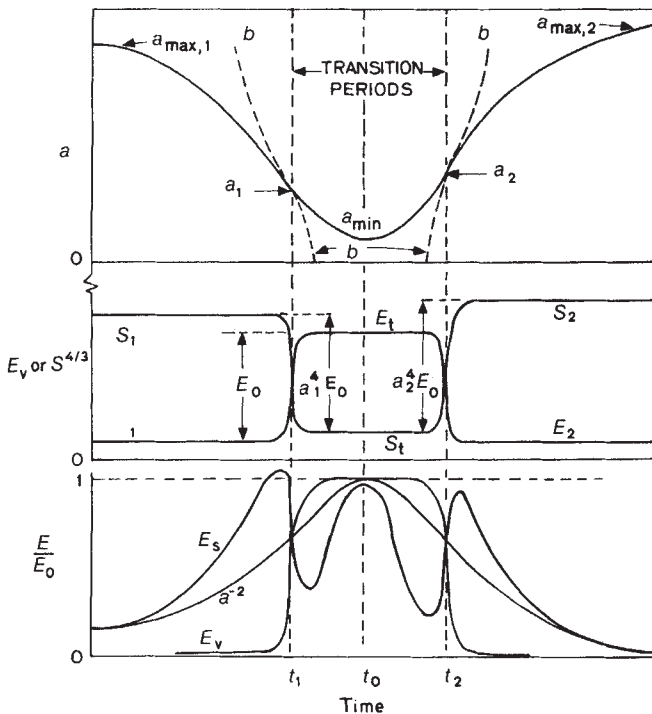


Fig. 2 Schematic variation with time of the expansion parameter (a), vacuum energy density (E_v), photon number (S), substance energy density (E_s), and the curvature term (a^{-2}), during a bounce. At t_1 and t_2 models *b* and *b'* (shown in Fig. 1) with two different values of S and with $a_{\min} < a_{\max}$ are joined smoothly. The minimum value of E_s at t_1 depends on the rate of the transition. The slower this rate the smaller the depth of the minimum. The minimum near t_2 depends on the degree of supercooling. Note that at $a_{\max}$, $E_s \approx a^{-2}$ and at $a_{\min}$, $E_v \approx a^{-2}$.

that the 10^{90} photons we observe today may have been produced through many successive bounces of the universe (see Fig. 3). After each bounce the number of photons increases as $S_{i+1}^{4/3} = S_i^{4/3} + E_0(a_{2,i}^4 - a_{1,i}^4)$, and the $\dot{a}^2 = 0$ curve in Fig. 1 moves to the right and decreases in amplitude. If the residual vacuum energy density $E_{v,i+1}$ is not zero, eventually as S increases, one may reach the condition $S_i > (4E_{v,i})^{-3/4}$ which will give rise to an ever-expanding Lemaitre type model as shown by model *c'* and by the heavy dashed line directed to the right in Fig. 1. However, note that as the symmetry is restored before each bounce, there must be a fine tuning of the exact epoch of the transition a_1 so that S_i , the equivalent number of photons and pairs during the transition period, is not too large; $S_i^{4/3} < 1/(4E_0)$. Otherwise, as discussed above, a bounce will not occur.

Discussion

Whether the above scenario will work in our Universe depends on the validity of the many assumptions involved. I now comment on these assumptions.

(1) The existence of the bounce requires an extremely small value for the total substance energy of the universe (more exactly the quantity $S_i^{4/3} = E_s a^4$) during the transition phase. We need $S_i < E_0^{-3/4}$ ($\sim 10^{15}$ for GUT) as compared with $S_i \sim 10^{90}$ for our Universe. This requires that, at the transition, E_s is very nearly equal to E_0 and that the transition occurs quickly so that the PdV work done by gravity during the transition is negligible. Such an extreme draining of energy from the substance clearly indicates that the two phases are not in thermal equilibrium or pressure balance. If equilibrium were maintained (say through a slow transition), then the substance energy would remain dominant (A. Guth, personal communication) and the entropy would be conserved, which, as shown by Sher¹⁵, precludes any possibility of bounce.

(2) We have also assumed that the total energy density and pressure of the universe can be separated into two constituent parts, one for the radiation with $P_s = 1/3 E_s$ and one for vacuum with $P_v = -E_v$. This clearly is true during times sufficiently far removed from the transition. Such a separation is also possible during the transition since, in general, given the total energy density E and pressure P with an arbitrary equation of state, we can always define $E_s = 3P_s = 3(E + P)/4$ and $E_v = -P_v = (E - 3P)/4$, so that $E = E_s + E_v$ and $P = P_s + P_v$. However, the important requirement of the bounce is that during the transition E_v must increase monotonically from its initial value E_1 to the final value $E_1 + E_0$. If this variation is not monotonic, that is, if during some part of the transition $\dot{E}_v < 0$, then, according to equation (4), S_i will not be as small as required and it may even exceed S_1 . Whether $\dot{E}_v$ remains positive throughout the transition depends on the details of the microphysics which determine the equation of state $P = P(E)$. We have shown that for the most commonly used (see, for example refs 1, 15) effective potential (or free energy) forms the vacuum part of the equation of state does, in fact, increase slowly and monotonically up to the temperature where the non-zero minimum of the potential disappears. As the temperature exceeds this value, the transition will probably occur very quickly. This will raise the vacuum energy density to its final value and, as demanded by energy conservation, will drain the substance of its energy.

Such a transition will destroy the thermal equilibrium of the substance and the normal relationship between entropy and energy density will no longer be valid. Thus, a decrease in the energy density E_s does not necessarily mean a decrease in entropy. The details of the variation of E_v and E_s and entropy will be discussed elsewhere.

(3) The most attractive feature of this scenario is that it allows for creation of substance through repeated bounces. This energy comes from the vacuum which, because of the nature of its equation of state ($P_v = -E_v$), can be the source for creation of an unlimited amount of energy as the universe expands: $dE = P_v dV > 0$. However, our scenario cannot be extended to very small values of a number of particles. It will work as long as S

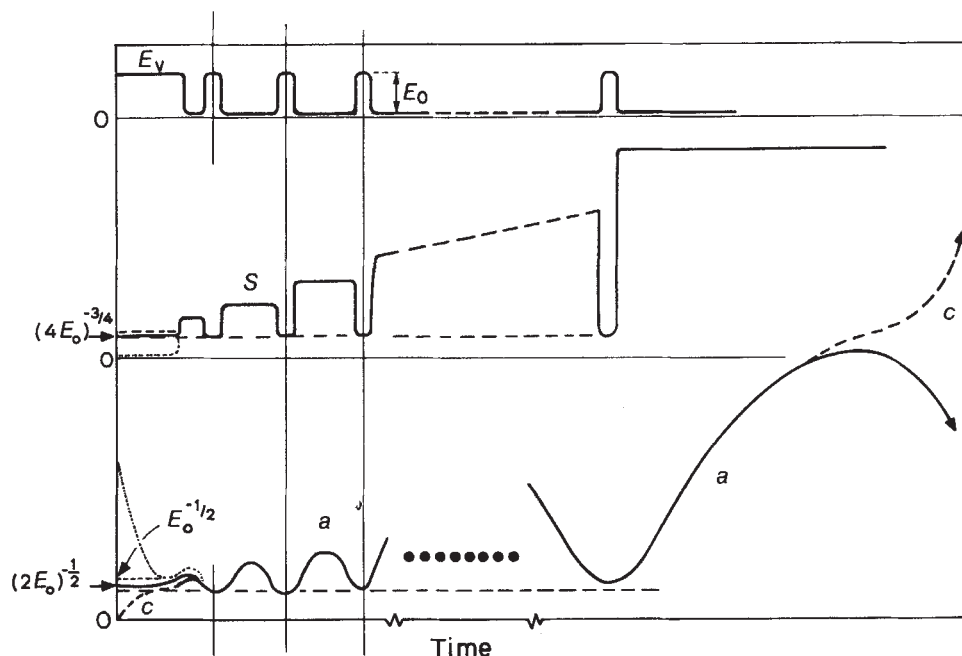


Fig. 3 Schematic variation of vacuum energy density (E_v), photon number (S) and expansion parameter (a) with time for repeated bounces. The oscillations stop if during the normal phase $4S_0^{4/3}E_0 < 1$; solid line for $x_0 = 1$ also Fig. 1.) Three possible initial conditions are shown. Dotted line is for initial $S_0 = 0$ (or $x_0 = 4S_0^{4/3}E_0 < 1$); solid line for $x_0 = 1$ has a static Einstein beginning; and the dashed line marked c (see also Fig. 1), with $x_0 > 1$, starts from a Planck era and undergoes a Lemaitre phase before the first phase transition.

in the symmetry broken (that is, normal) phase exceeds $(4E_0)^{-3/4}$ ($\sim 10^{15}$ for GUT). If S is less than this value, then $a_{\max} \sim S^{2/3}$ becomes less than $a_{\min} \sim E_0^{-1/2}$, indicating the breakdown of our analysis.

(4) There are various possible ways out of this dilemma with regard to the initial conditions. The universe may have started as a closed, vacuum-dominated deSitter universe, contracting from infinity to $a_{\min} = E_0^{-1/2}$ and then expanding with $a = +(E_0 a^2 - 1)^{1/2}$, as shown by the dotted lines in Fig. 3. If at $a = a_0$ the vacuum breaks down¹, then substance energy, with $S^{4/3} = E_0 a_0^4 > E_0^{-1}$ will be created. The expansion will continue at a new rate, $a = +(E_0 a^4/a^2 - 1)^{1/2}$, reaching a maximum expansion $a_{\max} = a_0^2/a_{\min}$. After this our previous scenario applies, the universe contracts, undergoes the usual symmetry restoring transition, rebounds at a new $a_{\min} > E_0^{-1/2}$, breaks the symmetry at a_2 and emerges with a higher value of S and so on. Another possibility is that the universe starts with $E_v = E_0$ and with some finite initial S_0 . If $S_0 < (4E_0)^{-3/4}$, then the universe starts as described above with $a_{\min} < (E_0)^{-1/2}$ [or as an Einstein static universe with $a_{\min} = (2E_0)^{-1/2}$ if $S_0 = (4E_0)^{-3/4}$, solid line beginning in Fig. 3]. The result here will not be much different from that of the deSitter beginning described above. However, if $S_0 > (4E_0)^{-3/4}$ then the universe must begin from a Planck stage and undergo the first phase transition after it has cooled down below T_0 (dashed line marked c in Fig. 3).

Therefore, if the vacuum does not spontaneously break down to create $S_0 > (4E_0)^{-3/4}$ pairs at once, then we must postulate a universe starting not only with $E_v = E_0$ but also containing $S_0 > (4E_0)^{-3/4}$ photons and pairs. We can relax this assumption if there exist other phase transitions at higher temperatures, with higher values of E_0 and, therefore, requiring lower values of S_0 . In particular, phase transition near the Planck era ($E_0 = 1$ and $S_0 = 1$) will allow the universe to start with very little substance and create most of it through successive bounces.

(5) Finally let us consider the so-called monopole, horizon and flatness difficulties^{16,17}. If the supercooling is extreme, the

expected density of monopole would be acceptable as in Guth's¹⁰ inflationary scenario. Furthermore, according to our scenario during the inflationary phase ($t_0 < t_2$) there are only $S_i (\sim 10^{15}$ for GUT) photons and particles in the whole universe. Thus some of the restrictions imposed on the conventional models by the monopole problem will be alleviated. For example, if one bubble is nucleated per particle, then the total number of monopoles in the present universe would be $< 10^{15}$ and therefore insignificant.

It can also be shown that during the transition phase of each bounce ($t_1 < t < t_2$) signals moving with speed of light can circumnavigate the whole universe eliminating the horizon problem. The flatness problem is no longer a difficulty because during both reversals of each cycle the curvature term is important. During the in-between intervals ($t < t_1$ or $t > t_2$) the time duration for near flatness becomes longer and longer as the entropy increases through repeated bounces.

(6) A bounce will occur if the symmetry restoration process induces an increase in the vacuum energy density and a corresponding decrease in the energy of substance. These changes do not have to be equal to those expected from the reverse transition. However, if the two processes are not nearly symmetric, the occurrence of repeated bounces could not take place as simply as described here and the fine tuning process needed for continuation of the bounces would break down. This would cause the universe to expand indefinitely (or collapse and reach temperatures above the critical temperature T_0 of the phase transition) prematurely; that is before it has built up 10^{90} photons (and 10^{80} baryons). Such a model will not resemble our Universe and can be disregarded by appeal to the anthropic principle¹⁸.

I thank Drs Lawrence Caroff and Robert Wagoner for critical readings of this manuscript and for valuable comments, also Drs Alan Guth, Marc Sher, Gary Steigman, Leonard Susskind and Michael Turner for interesting discussions. This work was supported in part by NASA-Ames Research Center grant NCA2-OR745-008.

Received 4 January; accepted 9 July 1982.

1. Linde, A. D. *Rep. Prog. Phys.* **42**, 390 (1979).
2. Lemaitre, G. E. in *Albert Einstein, Philosopher-Scientist* (ed. Schlegel, P. A.) 39 (Cambridge University Press, London, 1949).
3. Eddington, A. *Sci. Prog.* **34**, 225 (1939).
4. Zeldovic, Ya. B. *Soviet Phys. Usp.* **11** (1968).
5. Petrosian, V. in *Confrontation of Cosmological Theories with Observational Data* (ed. Longair, M. S.) 173 (Reidel, Dordrecht, 1974).
6. Wagoner, R. V. *Phys. Rev.* (1969).
7. Kazanos, D. *Astrophys. J. Lett.* **241**, L59 (1980).
8. Petrosian, V. & Salpeter, E. E. *Astrophys. J.* **151**, 411 (1968).

9. Gunn, J. E. & Tinsley, B. M. *Nature* **257** (1975).
10. Guth, A. *Phys. Rev. D* **23**, 347 (1981).
11. Linde, A. D. *Phys. Lett.* **108B**, 389 (1982).
12. Albrecht, A., Steinhardt, P. J., Turner, M. S. & Wilczek, F. Preprint no. 82.09 (Enrico Fermi Institute, 1982).
13. Bludmann, S. *Pap. at Workshop at Institute for Theoretical Physics, Santa Barbara* (1981).
14. Barrow, J. D. & Turner, M. S. *Nature* **298**, 801-805 (1982).
15. Sher, M. A. *Phys. Rev. D* **22**, 2989 (1980).
16. Preskill, J. P. *Phys. Rev. Lett.* **43**, 1365 (1979).
17. Turner, M. S. *Proc. 2nd Workshop on Grand Unification, Ann Arbor* (1981).
18. Carter, B. *IAU Symp.* No. 63, 291 (1974).

Trace elements in orogenic lherzolites reveal the complex history of the upper mantle

Michel Loubet & Claude J. Allègre

Laboratoire de Géochimie-Cosmochimie, Institut de Physique du Globe et Département des Sciences de la Terre, Université Pierre et Marie Curie, 4 Place Jussieu, 75230 Paris Cedex 05, France

Major and trace elements of several lherzolite-type bodies—at Lherz-Freychinède (French Pyrénées), Beni-Bouchera (Morocco), Lanzo (Italian Alps)—have been investigated. Quantitative model analysis of the element distributions shows that these bodies have undergone a rather complex series of events including successive partial meltings and subsequent differentiation of the melts produced. These events are related to the geodynamic evolution of these bodies within the upper mantle, and thus illustrate the multi-stage history of mantle material.

THE orogenic lherzolite type massifs¹, also called high-temperature peridotites² are thought to represent pieces of upper mantle material which were mechanically emplaced into the continental crust through large tectonic movements³⁻⁵. These massifs typically outcrop along major fault systems within chains where interactions between oceanic and crustal materials have occurred (for example, the Insubric line for the Alps massifs such as Lanzo and Baldissero and the North Pyrenean fault for the Pyrenean massifs such as Lherz and Freychinède).

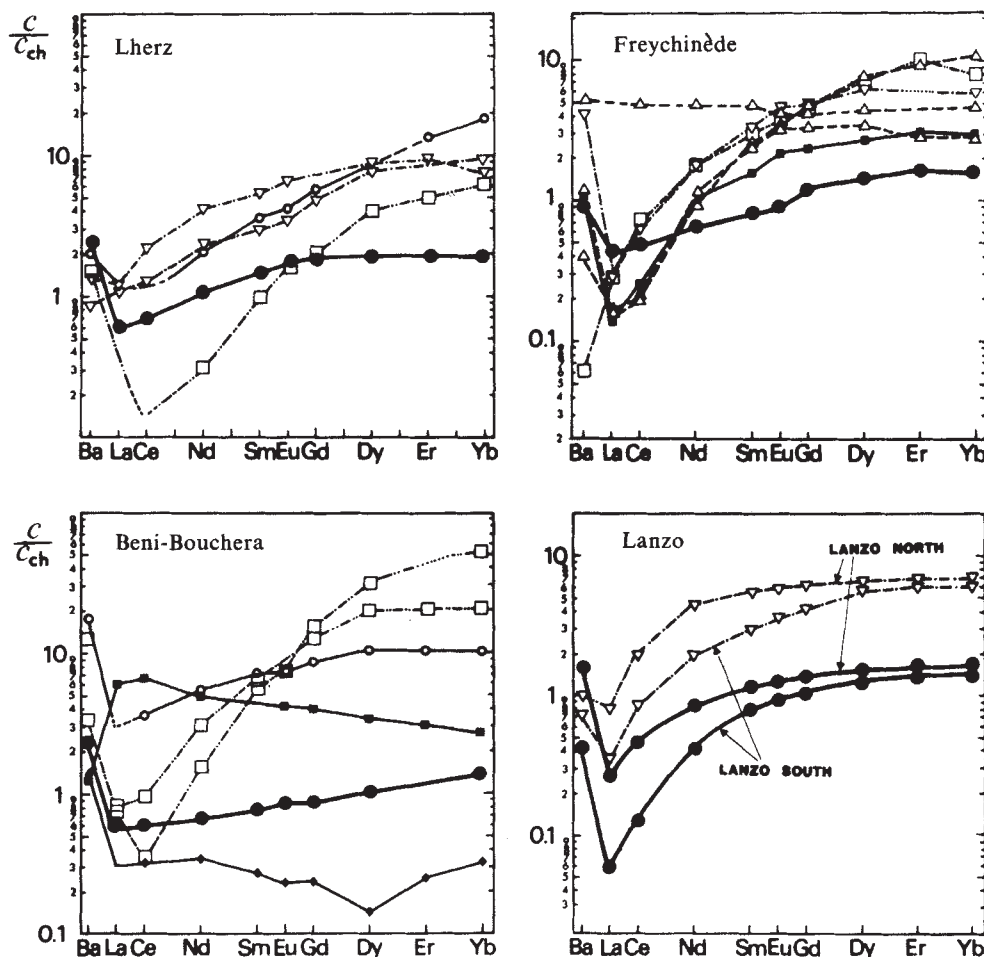
These massifs are therefore unique and highly interesting because they occur as large bodies relative to the size of xenoliths embedded in highly reactive magmas, and are therefore exposed to possible contamination by their surroundings. Oxygen isotope studies show that the orogenic lherzolite type

massifs are fresh representative samples of the upper mantle⁶ and their analysis should provide important information on the composition, the structure and the evolution of the Earth's upper mantle.

Petrological and geochemical structure of the massifs

Many petrological and geochemical studies have already been performed on these massifs and, in particular, on Mt Albert (Canada)⁷, Lizard (England)⁸, Beni-Bouchera (Morocco)^{9,10}, Ronda (Spain)^{10,11}, Lherz (France)¹², Lanzo (Italian Alps)^{13,14} and various Alps¹⁵ massifs. Geochemical analyses were carried out at the Lizard (England)^{16,17}, Lherz and various Alps¹⁸⁻²⁴,

Fig. 1 Chondritic normalized representation of the REE + Ba content of some pyroxenite layers from the orogenic lherzolite bodies studied. Some 'common' lherzolites from these bodies have been plotted for reference. ●, Massive lherzolites; ■, olivine websterites; ◆, enstatites; ○, complex layer; △---△, garnet websterites; ▽---▽, garnet free websterites; □, garnet clinopyroxenites.



Beni-Bouchera (Morocco)²³⁻²⁶, Ronda (Spain)^{27,28}, Liguria (Italy)²⁹ bodies.

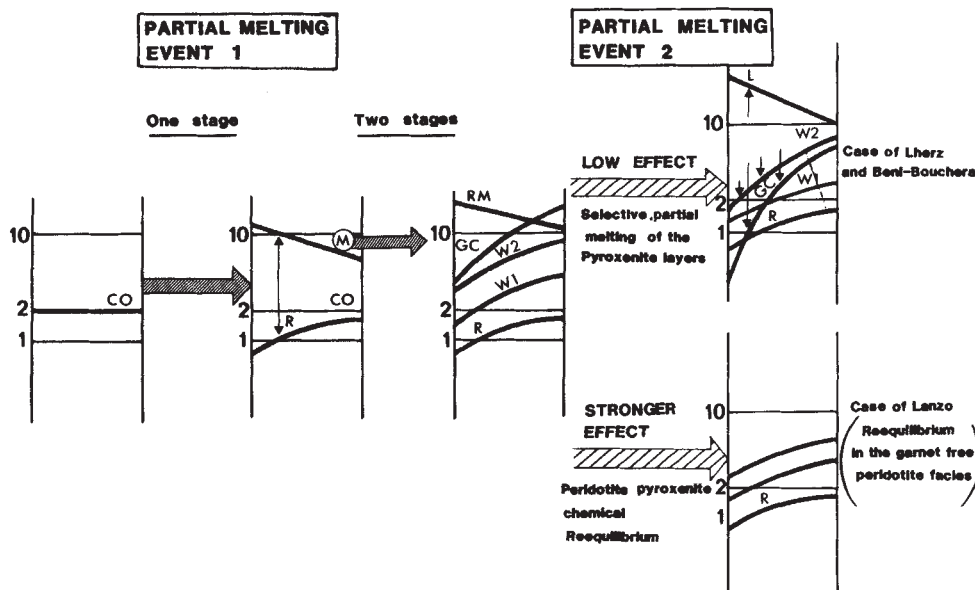
We have studied in detail the following lherzolite type orogenic bodies: Lherz and Freychinède, French Pyrénées (see ref. 12 for petrographic descriptions); Beni-Bouchera, Morocco (see refs 9, 10 for petrographic descriptions); Lanzo, Italian Alps (see refs 13, 14 for petrographic and structural descriptions). In other massifs from the Alps—Locana, Rivara, Baldissero and Alpe Arami¹⁵—only the lherzolites were looked for. We report here our results from the trace-element viewpoint and also present our main data. Detailed studies will be published elsewhere.

Orogenic lherzolite-type massifs are mainly formed of apparently rather homogeneous lherzolite which constitutes around 90–95% of the whole bodies. The remaining part is formed by mafic layers of various sizes apparently more or less parallel to one another. These layers are chemically basaltic but mineralogically variable, their modal composition extending from that of olivine websterite to that of garnet clinopyroxenite (pyroxenite layers). In some of the massifs late amphibole-bearing-type or gabbroic-type dykes cut across the mafic layers. Our data, presented in Tables 1 and 2 with details on the experimental procedure, include major element, rare-earth elements (REE) and transition element contents. The REE data

Table 1 Rare-earth element data from various peridotite bodies

	Ba	La	Ce	Nd	Sm	Eu	Gd	Dy	Er	Yb	Lu
Beni-Bouchera											
Peridotites											
M5-102	1.03	0.20	0.61	0.59	0.236	0.0985	0.386	0.537	0.342	0.337	
BB5	0.40	0.145	0.52	0.65	0.257	0.110	0.423	0.584	0.383	0.40	
M6-128	8.64	0.17	0.48	0.382	0.14	0.062	0.218	0.313		0.27	
Pyroxenites											
M6-79	45.5	0.22	0.29	1.24	1.03	0.55	3.89	9.41		10.0	
M7-58	63.3		2.90	3.20	1.31	0.534	2.23	3.20	1.89	1.93	
M5-367	12.1	0.24	0.75	1.78	1.26	0.641	3.35	6.09	3.78	3.99	
M7-701	1.11		0.30	0.35	0.12	0.056	0.178	0.21	0.148	0.16	0.03
M7-703	10.0	1.1	3.8	3.76	1.63	0.704	2.53	4.65		2.98	
M7-702	13.3	0.75	2.6	3.52	1.49	0.64	2.61	3.87	2.41		
M5-103	4.5	1.85	5.23	3.20	0.30	1.05	1.05	1.05	0.54	0.50	0.052
M6-78	5.2		0.25	0.20	0.050	0.016	0.061	0.43	0.45	0.60	0.12
Alps											
Peridotites											
Lanzo											
L705b	1.49	0.018	0.103	0.271	0.144	0.0656	0.271	0.398	0.270	0.29	
775	1.03	0.067	0.36	0.537	0.225	0.0954	0.375	0.492	0.309	0.327	
C50	0.323	0.0506	0.029	0.271	0.131	0.057	0.269	0.355	0.224	0.258	
Other Alps massifs											
Locana: L4	4.25	0.12	0.464	0.461	0.172	0.074	0.272	0.383	0.242	0.248	
Rivara: R1	1.72	0.027	0.124	0.278	0.144	0.067	0.326	0.407	0.263	0.285	
Baldissero: Ba	0.364	0.028	0.101	0.165	0.105	0.0484	0.216	0.345	0.240	0.266	
Alpe Arami: AA	1.02	0.167	0.623	0.563	0.230	0.100	0.433	0.698	0.460	0.484	
Pyroxenites											
Lanzo body											
779-b	3.6	0.24	1.66	2.61	0.96	0.38	1.53	2.03	1.23	1.26	
741-b	2.56	0.092	0.66	1.15	0.58	0.26	0.11	1.70	1.05		
Gabbro dykes and associated peridotites											
Lanzo											
751-3	20.8	0.48	0.90	0.51	0.124	0.246	0.17	0.17		0.0073	
C-55	5.52	0.46	0.87	0.62	0.17	0.236	0.24	0.284		0.13	
L684-d	12.7		1.66	0.82	0.82	0.58	1.41	1.77	1.02	0.87	
751-1G	7.3		0.03	0.014	0.0032	0.0097					
751-2G	0.61	0.028	0.093	0.112	0.052	0.024	0.090	0.122	0.073	0.082	
Freychinède-Lherz											
Freychinède peridotites											
71-335	3.3	0.12	0.376	0.40	0.15	0.0654	0.32	0.424	0.30	0.30	
Lherz peridotites											
70-367	8.6	0.16	0.54	0.66	0.265	0.11	0.48	0.55		0.34	
LZM1		0.16	0.50	0.483	0.322	0.116	0.39	0.483		0.322	
71-325	0.79	0.013	0.036	0.039	0.011	0.0049	0.019	0.020		0.011	
Freychinède pyroxenites											
74-57	17.7		4.03	2.8	0.98	0.30	1.35	1.37		0.9	
72-50	1.44	0.051	0.16	0.56	0.465	0.24	1.22	2.33	1.68	1.95	
70-357	4.29	0.045	0.168	0.64	0.45	0.234	1.22	2.33	1.68	1.95	
72-411	3.76	0.04	0.21	0.61	0.286	0.15	0.604	0.83	0.58	0.53	
72-49	14.7	0.084	0.51	1.00	0.60	0.33	1.21	1.84		1.05	
72-224	0.22	0.089	0.55	1.6	0.56	0.274	1.24	2.29	1.72	1.52	
Lherz pyroxenites											
70-379	7.2			1.22	0.66	0.30	1.05		2.47	3.4	0.55
70-291	5.6			0.18	0.180	0.117	0.51	1.23	0.91	1.14	
72-445	3.12	0.33	1.01	1.30	0.55	0.25	1.23	2.30		1.86	
69-356W	51.3	0.35	1.76	2.77	1.10	0.46		2.72	1.69	1.42	
69-356	8.3		0.84	0.33	0.136	0.136	0.524	0.71	0.45	0.47	0.075
73-78W	3.62	9.8	11.1	4.6	0.754	0.294	0.77	0.92	0.55	0.61	
73-78	1.88	0.716	2.04	0.954	0.157	0.0636	0.179	0.200	0.140	0.152	
Minerals											
70-148 diopside	1.27	0.22	0.26	0.42	0.29	0.16	0.53	0.44		0.10	
70-367 diopside	501	0.88	2.66	3.35	1.38	0.56	2.21	3.10		2.14	

Fig. 2 Summary of the REE content evolution which, according to our analysis, has affected the various products found in the lherzolite-type bodies—in particular, the pyroxenite layers—during the various episodes of their history. The REE patterns are represented by the classical chondrite normalized diagram. Step 1: the pyroxenite layers are created through partial melting of a peridotitic upper mantle (CO), either undepleted (as in this figure) or depleted. This partial melting yields (one stage) residual peridotites (R) and melts (M) which can differentiate (two stages) through fractional crystallization into different cumulate types: websterites (types W₂ or W₁) or garnet pyroxenites (GC) and residual melts (RM). Step 2: a more recent partial melting event affects these pyroxenite layers. The effect of this last event may be rather small, affecting the pyroxenite layers only selectively—(thus we explain the Lherz-Freychinède and Beni-Bouchera features) or stronger and, then, affecting also the peridotites. In the latter case, a general reequilibration of the pyroxenite layers with the peridotites will occur (the Lanzo features are thus explained).



are shown in Fig. 1 according to the classical chondritic normalized diagram.

For data analysis we will consider in turn the common lherzolites, the pyroxenite layers and the late dykes. We have used quantitative models based on the concept of partition coefficient and on the formulation of partial melting and fractional crystallization processes for trace elements. For each case experimental data are compared with the results of computed models.

Lherzolites

Although apparently homogeneous, those peridotites that constitute the major part of the bodies exhibit significant variations in composition which can even exist in a single massif within several hundreds of metres of each other.

Major elements—some correlations are observed between CaO, Al₂O₃, Na₂O and TiO₂ as well as between MgO, NiO, Cr₂O₃ contents and the MgO/FeO ratio^{9,10,18}.

Trace elements—the distribution observed can be summarized as follows^{20–27}: as far as we know, the elements characterized by high partition coefficients (highly compatible elements (Ni, Co and to a lesser extent Cr)) have a remarkably homogeneous content. Elements with intermediate partition coefficients such as Yb and Ti have content variations which do correlate with the CaO variations. Large variations are observed for the incompatible trace elements. In particular, the REE patterns of the lherzolites show remarkably similar characteristics. They are all light REE depleted, the light REE contents varying within a rather wide range and the heavy REE contents in a rather narrow one^{18–20}. Early results confirmed the preliminary study by Haskin *et al.*³⁰ and Frey^{16,17} and were later supported by further data^{27–29}.

In agreement with Frey¹⁶, we interpreted¹⁸ these characteristics as those of a residual material which originated through partial melting of a rather homogeneous primitive peridotitic mantle. This primitive upper mantle may have been characterized by a REE chondrite normalized pattern parallel to the chondritic line of reference but with about double the chondritic value¹⁸. The most important conclusions reached from the quantitative models developed^{18,19,26,27,28} are that the residual peridotites resulting from partial melting of an originally homogeneous lherzolitic upper mantle display a wide range of variations in light REE contents and a relatively small range of variations for the heavy REE—a feature effectively observed in the orogenic lherzolites.

The degree of partial melting which gave rise to these residual peridotites can be estimated from REE analysis to vary between 1 and 5% (ref. 18). The state of equilibrium of these peridotites during their partial melting is unclear due to uncertainty over the REE partition coefficients for clinopyroxene. This residual character also explains the small variation of Ni, Co, Cr and the covariation of CaO, Al₂O₃, Na₂O, TiO₂. Nd and Sr isotopic data corroborate the residual nature of this peridotitic upper mantle material^{22–24}. Their Nd isotopic composition moves along the same Nd–Sr anticorrelation line as the one characteristic of the oceanic basalts, especially in the zone corresponding to high Nd isotopic values²⁴. At the same time Nd isotopic data exclude alteration as a possible process for light REE depletion of these peridotites—as proposed elsewhere²⁹.

Thus, major elements, trace elements and isotope data agree with the idea that orogenic lherzolites are residues after extraction of a small amount of basaltic melt. This interpretation is supported by most authors who have studied these massifs^{24,25,27–29}, and agrees with the general idea, based on oceanic basalt trace elements^{31–33} and isotope analysis^{34–36}, that the 'oceanic' upper mantle should be residual (extraction of small basaltic amount). At the same time this agreement suggests that the orogenic lherzolite massifs could effectively represent oceanic upper mantle material.

Pyroxenite layers

The chemical composition of these layers as a whole is basaltic. Thus, one could think of these layers as representing the peridotite complement, that is the part which was extracted during the partial melting of these rocks. However, the detailed mineralogical and chemical analysis of these layers implies that the process cannot be so simple. In particular it looks as though some differentiation processes should have occurred, as indicated by the symmetrical structure displayed by some complex pyroxenite layers formed by the superposition of several sub-layers^{9–12}. The models proposed for the origin of these layers should thus take this complexity into account.

REE contents of the pyroxenite layers

The REE patterns of the large set of pyroxenite layers is illustrated in Fig. 1 through the classical chondrite normalized representation. For the various lherzolite type bodies we studied, the main characteristics of these patterns are:

(1) In general the pyroxenite layers look more REE-enriched

Table 2 Major and transition elements in various peridotite bodies

	Beni-Bouchera					Freychinède					
	Peridotites		Pyroxenites			Peridotite	Pyroxenites				
	BB5	M5-102	M5-367	M6-79	M7-58	71-335	72-211	72-50	72-49	70-357	74-57
Major elements											
SiO ₂	41.82	42.73	43.10	44.67	44.29	43.29	43.00	47.55	46.11	46.65	54.40
Al ₂ O ₃	3.50	3.29	14.76	112.42	10.61	2.65	5.37	9.75	11.40	12.06	13.77
Fe ₂ O ₃	9.06	9.19	17.35	13.96	10.65	9.16	11.20	8.51	8.24	8.18	8.39
MgO	37.99	38.90	11.20	13.87	17.50	40.32	0.12	0.18		0.18	
MnO					0.19	0.13	33.41	21.47	23.60	18.48	19.26
CaO	3.02	3.00	10.66	11.56	11.85	2.63	4.25	10.49	8.95	12.49	11.90
Na ₂ O				1.24	0.96	0.24	0.08	0.82	0.66	0.52	0.67
K ₂ O	0.014	0.015	0.02	0.02	0.026	0.018	0.018	0.013	0.03	0.009	0.01
TiO ₂	0.14	0.14	1.01	0.50	0.55	0.14	0.232	0.32	0.46	0.405	0.38
P ₂ O ₅	0.01	0.02	0.05	0.04	0.037	0.016	0.008	0.018	0.05	0.024	0.03
P at 110 °C	0.20	0.09		0.11	0.20	0.08	0.12	0.05	0.09	0.05	0.07
P at 1,050 °C	3.87	1.68		0.84	2.03	0.95	1.03	0.07	0.74	0.05	0.14
Total						99.498	98.383	99.241		99.098	
Transition elements											
Ti	840	840	6,060	3,000	3,300	634	1,392	1,920	2,760	2,430	2,280
V	42		381		305	58	92	252	162	232	205
Cr	1,340		600		1,120	2,210	1,940	1,370	2,200	1,080	875
Mn					1,470	1,000	930	1,390		1,390	
Fe					74,490	64,570	78,340	59,520		57,210	
Co	103		90		105	115	115	86	53	85	66
Ni	1,940		1,550		1,480	2,820	2,450	1,720	1,550	2,240	600
Cu	41		7		506	25	61	66	68	74	90
Zn	85		129		68	64	60	50	56	34	36
				Lherz			Amphibole-bearing dykes				
				Pyroxenites							
Major elements	71-325	73-78	70-379	70-291	72-445	68-110	70-242				
SiO ₂	41.44	41.94	44.88	45.95	44.54	42.67	40.12				
Al ₂ O ₃	0.65	2.07	12.33	12.92	12.04	12.71	13.36				
Fe ₂ O ₃	8.71	8.66	8.39	7.91	6.35	12.13	10.37				
MgO	0.12	0.12	0.21	0.18	24.85	14.29	14.42				
MnO	44.99	40.78	20.96	16.30	24.85	14.29	14.42				
CaO	0.39	1.72	10.74	14.80	9.11	10.97	10.36				
Na ₂ O	0.02	0.15	0.40	0.66	0.58	1.71	2.75				
K ₂ O	0.014	0.015	0.022	0.013	0.01	0.56	1.40				
TiO ₂	0.015	0.058	0.194	0.100	0.29	2.74	5.06				
P ₂ O ₅	0.004	0.019	0.029	0.022	0.05	0.07	0.07				
P at 110 °C	0.15	0.33	0.05	0.069	0.19	0.21	0.16				
P at 1,050 °C	2.69	3.03	0.55	0.25	2.34	1.24	1.75				
Total	99.193	98.892	98.665	99.073							
Transition elements											
Ti	90	348	1,164	600	1,740	16,440	30,360				
V	25	50	350	257	189	540	198				
Cr	1,940	2,580	1,460	1,720	2,420						
Mn	930	930	1,630	1,390							
Fe	60,920	60,570	58,680	54,620							
Co	115	113	80	95	113	158	25				
Ni	2,840	2,880	2,250	2,700	1,600	1,500	500				
Cu	10	25	49	6	45	142	24				
Zn	58	64	52	46	87	128	32				
				Lanzo							
				Peridotites			Pyroxenites	Peridotites associated with the gabbro dykes			
Major elements	RI	Ba	AA	705B	775	779b	741b	751-1G	751-2G		
SiO ₂	41.68	42.92	41.90	42.82	42.31	48.29	48.61	36.89	42.68		
Al ₂ O ₃	2.66	2.71	3.40	2.64	3.05	7.17	6.39	0.69	1.63		
Fe ₂ O ₃	8.68	9.73	9.48	9.11	9.01	6.37	6.40	10.29	9.26		
MgO	38.42	41.34	39.62	40.42	38.80	0.12	0.12	0.12	0.12		
MnO	0.13	0.13	0.14	0.12	0.12	26.27	24.80	46.76	41.20		
CaO	2.90	2.64	3.21	2.84	3.02	8.87	11.46	0.14	2.32		
Na ₂ O	2.90	2.64	3.21	0.03	0.20	0.58	0.58	0.010	0.016		
K ₂ O	0.014	0.015	0.052	0.015	0.015	0.012	0.015	0.019	0.044		
TiO ₂	0.095	0.08	0.109	0.10	0.13	0.49	0.27	0.009	0.011		
P ₂ O ₅	0.012	0.012	0.017	0.008	0.015	0.020	0.025	0.23	0.08		
P at 110 °C	0.27	0.04	0.11	0.12	0.13	0.16	0.12	4.33	1.31		
P at 1,050 °C	3.91	0.38	1.07	0.31	2.37	0.69	0.68	99.538	98.861		
Total	98.87	99.26	99.23	98.533	99.17	99.042	99.510	114	264		
Transition elements											
Ti	570	480	654	600	780	2,940	1,620	114	264		
V	55	58	55	58	58	140	145	21	52		
Cr	2,440	2,480	2,930	2,820	2,680	3,280	3,030	3,680	3,920		
Mn	1,000	1,000	1,080	930	930	930	930	930	930		
Fe	60,710	68,050	66,300	63,718	63,020	44,550	44,760	71,970	64,770		
Co	125	122	120	150	125	120	100	132	140		
Ni	1,900	2,240	2,180	2,220	2,180	1,140	1,400	2,600	2,260		
Cu	26	28	28	30	76	26	125	12	39		
Zn	70	66	68	76	70	70	50	88	70		

Major elements were analysed by X-ray fluorescence, transition elements by atomic absorption, rare earths by isotope dilution mass spectrometry with an accuracy of better than 2.5%. The experimental techniques have been described elsewhere²⁰.

than common peridotites. In particular the heavy REE content of some pyroxenite layers appears very high, for example, that of garnet clinopyroxenite layers (M5-367, M6-79 complex layers).

(2) They display a large variety of REE patterns characterized by a wide range of light REE fractionations relative to the heavy REE, some being light-REE enriched (74-57 in Frey-chinède, M5-103 in Beni-Bouchera), others light-REE depleted.

(3) The light REE depleted pyroxenites are the most common with a tendency to be more light-REE depleted relative to the heavy REE than are the common lherzolites.

(4) No distinct Eu anomaly is observed in these patterns (with the exception of the websterite 73-78W from Lherz).

Quantitative models for estimating the REE contents that should represent the products resulting from partial melting of a lherzolitic upper mantle have been developed. The basic hypothesis is that the primitive upper mantle affected has a homogeneous and uniform composition.

In the single stage model we try to compute the composition of the various types of melts that are generated through a single partial melting. More complex models are needed to describe the composition of the various cumulates which may crystallize from these melts and the corresponding residual liquids after an evolution to these melts through fractional crystallization (two stage process). The most important conclusions of this quantitative modelling are: (1) the melts produced have a large diversity in their REE patterns, characterized by a wide range of fractionation of the light REE relative to the heavy REE. Taking into account the fractionated type partial melting process, these melts may indeed be light REE enriched or depleted relative to the heavy REE; (2) the cumulates which crystallize from these melts thus exhibit, as well as the melts, a wide range of REE patterns also characterized by a large variety of fractionations of the light REE relative to the heavy REE. This diversity in the REE patterns will be accentuated in the cumulates, given the variety of mineralogical assemblages which may crystallize; for example, the crystallization of garnet in the cumulates will yield a strong enrichment of such cumulates in heavy REE.

The large diversity of the REE patterns displayed by the pyroxenite layers conforms with what would be expected for products derived through partial melting of a peridotitic upper mantle, according to the models developed. However, we cannot decide from these REE data, whether most of the layers are cumulates or melts, because of the large range of contents these two types of products can display in such a model. To solve this problem, the use of elements with high partition coefficients appears adequate. Cumulates and melts should have very different properties for this kind of element.

Trace elements with high partition coefficients

Large variations are found in the literature for the partition coefficients of this kind of element so that the models obtained may be rather imprecise. A correct estimate may, however, be realized in a special case, that is for the cumulates supposed to correspond to the first per cent of crystallization of the primary melts. The phases of these first cumulates should have the same trace-element content as the phases of the common peridotites. So, taking values for peridotite phases from the literature, we estimated the content of several types of cumulates presenting different modal compositions, all supposed to have crystallized during the first per cent of crystallization of primary melts. This gives us some kind of upper limit for the content of the cumulates²⁶.

The pyroxenite layers seem to contain a rather large amount of such elements (Ni, Co, Cr), compared with our estimates. This convinces us that most of the pyroxenite layers should be cumulates representatives. Possible exceptions to this rule include some complex layers, the contents of which are close to melt estimates²⁶.

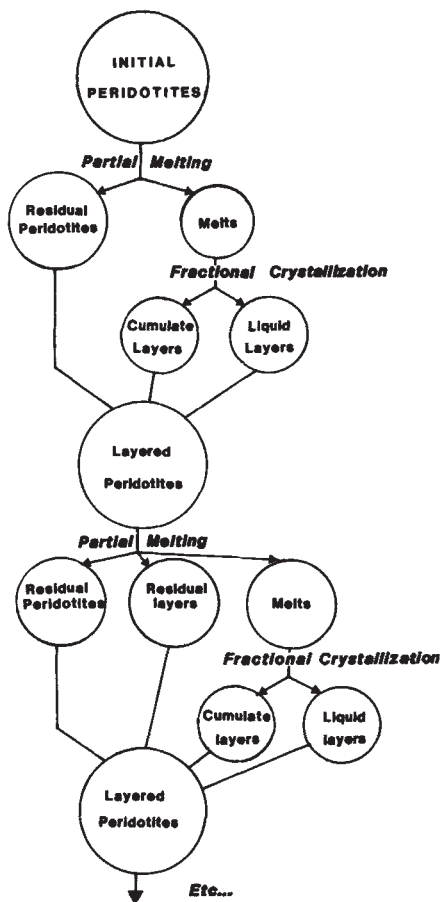


Fig. 3 Deduced model of evolution of the lherzolite type bodies. These peridotite bodies have been the object of successive melting events, each one producing melts possibly mobile and which differentiate through fractional crystallization. These derived products yield the pyroxenite layers. A later partial melting of this layered peridotite affects the various materials with variable intensities. Other types of products and layers will be generated through this new partial melting event.

Evidence for remelting

However, several observations suggest that most of the pyroxenite layers do not present primary structures and have been modified through a further process. More precisely we suspect that these pyroxenite layers have been affected after their formation by a further selective late partial melting event¹⁹⁻²².

The main points supporting this hypothesis are first, in several bodies the rocks which have the lowest content of highly incompatible elements are the garnet clinopyroxenite layers. Such are the garnet clinopyroxenites 70-291 from Lherz, and M6-79 from Beni-Bouchera which have the lowest light REE content, even lower than the common lherzolites from these bodies. Such a low light REE content is surprising in these clinopyroxene-rich rocks which represent the latest stage of differentiation of the melts yielding the pyroxenite layers.

Second, in some peridotite bodies a peculiar evolution is observed³⁷: a depletion in the content of the highly incompatible trace elements characterized by low partition coefficients when going from the slightly differentiated to the highly differentiated layers (from the olivine websterites to the garnet clinopyroxenites).

Third, some of the lherzolite type bodies that we studied effectively bear the mark of such a late partial melting event in the form of late dykes cutting the websteritic layering—in particular the amphibole bearing dykes—as in Lherz and Frey-chinède (French Pyrénées). The geochemical analysis of these dykes shows^{20,37} that they may effectively represent products

derived from the melts resulting from partial melting of these pyroxenite layers.

Sr isotope studies performed on these same rocks support this interpretation of a complex history of these layers²³ and were the basis for the Rb–Sr systematics given by Polve and Allègre²³.

The occurrence of such a late partial melting event has a direct consequence. As the primary trace element distribution of the pyroxenite layers has changed, a precise interpretation of the characteristics of the melts at the origin of these layers becomes difficult.

In the framework of our model for the origin of the pyroxenite layers from melts resulting from partial melting of the peridotitic mantle we would have liked to define precisely the nature of these melts. In Fig. 2, we retrace in the chondrite normalized representation, the various steps which according to our interpretation explain the actual REE content of these layers. They correspond to the different partial melting events suffered by these bodies.

Late dykes

In the Lherz massif late amphibole-bearing dykes do exist which cut across the pyroxenite layering³⁸. The REE patterns of these dykes are characteristic—all being enriched in light REE relative to the heavy REE. From their trace-element distributions we interpret these dykes as representative of cumulates crystallized from melts highly enriched and fractionated in light REE relative to the heavy REE. We have already supported the idea that these melts may have been issued from the partial melting of the pyroxenite layers^{20–37}.

In the Lanzo body, the REE patterns displayed by the gabbroic-type dykes is characteristic of plagioclase bearing cumulates. In this massif there are some indications that the last partial melting event at the origin of these dykes also affected the peridotites^{13,14}.

Conclusions

The analysis of these orogenic lherzolites yields strong evidence that several magmatic events affected these bodies. One was at the origin of the formation of the pyroxenite layers. The last one affected the pyroxenite layers in a selective way in some massifs (Lherz and Freychinède–Beni-Bouchera) but also reached the peridotites in others (Lanzo). Other partial melting events may have previously affected these lherzolite bodies thereby giving their residual character to the peridotites. Each magmatic event is characterized by the formation of various melt types through partial melting, a possible mobility of the melts produced and further fractional crystallization of the melts.

Thus, the final image which emerges is that the actual state of the orogenic lherzolite bodies can be understood through a

model in which these peridotite bodies suffered a complex history involving successive partial meltings and further differentiation of the melts so generated, the partial melting events giving rise each time to a diversity of products. The ⁸⁷Rb–⁸⁷Sr isotope systematics has also been interpreted in the light of a complex process²³. Dickey *et al.*²⁸, through the study of the Ronda massif (Spain), also support a similar complex evolution for this body but without clear evidence for pyroxenite-layer remelting. Wilshire and Shervais³⁹ inferred that the upper mantle parts should have suffered a rather complex evolution through a petrological and structural analysis of ultramafic xenoliths from the western United States.

Such a complex history can be understood geodynamically as resulting from an evolution of these bodies in an upper mantle affected by convection processes and uprising movements to the surface through tectonic events. The partial melting events thus appear during the decompression stages^{22,23,26}.

Several characters displayed by these Alpine-type peridotite bodies can be explained in the light of such an evolution. First, the relatively uniform lherzolites characteristics found throughout these bodies would result from an evolution of these upper mantle bodies in active convection processes, implying large mixing and chemical reequilibration—favoured by the presence of melts and fluids generated through the partial melting events. Second, the specific characteristics of the pyroxenite layers in each body may be explained by the relative intensity of the last partial melting event (the one which selectively affected these layers) (Fig. 3). In some bodies (Lherz, Freychinède, Beni-Bouchera) this last event was of low intensity. It affected only the pyroxenite layers in a selective way. In others, such as Lanzo, this last event was more intense, also affecting the peridotites^{13,14}. In this body the pyroxenite layers look perfectly equilibrated relative to the lherzolites.

This reequilibration should be a general phenomenon in the mantle—the temperatures being rather high, recrystallization occurring rather easily, the presence of melts or fluids permitting an efficient equilibration⁴⁰.

The apparent equilibrium of many websteritic layers with the surrounding peridotites in these bodies may result from such a further reequilibration process of ancient pyroxenite layers. In this case, these websteritic layers may be indicative of past partial melting events.

The oceanic mantle is composed of a mechanical mixture of an old subducted, elongated, sheared oceanic crust with a residual peridotite. Through several cycles of convection, including rising and melting at ridge crests, subduction with some rare melting events, origin of calcalkaline suites, and so on, we certainly have a quite complex petrological assemblage. Such a sequence of events, is in full agreement with petrological and isotopic data on orogenic lherzolite massifs^{23,24}. If this were verified, the pyroxenite would represent the old oceanic crust, sometimes recycled, or melted in subduction zones and the peridotites, the residual mantle.

Received 16 February; accepted 19 May 1982.

- Jackson, E. D. & Thayer, T. P. *24th Int. geol. Cong. Montréal*, Sect. 2, 289–296 (1972).
- Green, D. H. *Ultramafic and Related Rocks* 212–221 (Wiley, New York, 1967).
- MacKenzie, D. B. *Bull. Am. geol. Soc.* **71**, 303–318 (1960).
- Ave Lallemant, H. G. *Leid. geol. Meded.* **42**, 1–57 (1967).
- Nicolas, A., Bouchez, J. L. & Boudier, F. *Tectonophysics* **14**, 143–171 (1972).
- Javoy, M. *Colloques int. Cent. natn. Rech. scient.* **272**, 279–287 (1980).
- Smith, C. H. & MacGregor, I. D. *Bull. geol. Soc. Am.* **71**, 1978 (1960).
- Green, D. H. *J. Petrol.* **5**, 134–188 (1964).
- Kornprobst, J. *Contr. Miner. Petrol.* **23**, 283–322 (1969).
- Dickey, J. S. *Miner. Soc. Am. Spec. Pap.* **3**, 33–49 (1970).
- Obata, M. thesis, Massachusetts Inst. Technol. (1977).
- Conquére, F., *Bull. Soc. fr. Minér. Cristallogr.*, **100**, 42–80, 123–137 (1977).
- Boudier, F. thesis, Univ. Nantes (1976).
- Boudier, F., Nicolas, A. *Bull. suisse Miner. Petrol.* **52**(1), 39–58 (1972).
- O'Hara, M. J. & Mercy, E. L. P. *Earth planet. Sci. Lett.* **1**, 295–300 (1966).
- Frey, F. A. *Geochim. cosmochim. Acta* **33**, 1429–1447 (1969).
- Frey, F. A. *Phys. Earth planet. Inter.* **3**, 323–330 (1970).
- Loubet, M., Shimizu, N. & Allègre, C. J. *Contr. Miner. Petrol.* **53**, 1–12 (1975).
- Loubet, M., Bougault, H., Shimizu, N. & Allègre, C. J. *EOS* **67**, 1025 (1976).
- Loubet, M. thesis, Univ. Paris (1976).
- Loubet, M. & Allègre, C. J. *4th int. Conf. Geochronology Cosmochronology Isotope Geology*, 259–262 (1978).
- Loubet, M., Polvé, M., Richard, P. & Allègre, C. J. *Colloques int. Cent. natn. Rech. scient.* **272**, 269–277 (1980).

- Polvé, M. & Allègre, C. J. *Earth planet. Sci. Lett.* **51**, 71–93 (1980).
- Richard, P. & Allègre, C. J. *Earth planet. Sci. Lett.* **47**, 65–74 (1980).
- Menzies, M., Blanchard, D., Brannon, J. & Korotev, R. *Contr. Miner. Petrol.* **64**, 53–74 (1977).
- Loubet, M. & Allègre, C. J. *Geochim. J.* **13**, 69–75 (1979).
- Suen, C. J. thesis, Massachusetts Inst. Technol. (1978).
- Dickey, J. S., Obata, M. & Suen, C. J. *Bull. Oregon Dept. Geol. Min. Indust.* 79–89 (1977).
- Ottomello, G., Piccardo, G. B. & Ernst, W. G. *Geochim. cosmochim. Acta* **43**, 1273–1284 (1979).
- Haskin, L. A., Frey, F. A., Schmitt, R. A. & Smith, R. H. *Phys. Chem. Earth* **7**, 169–321 (1966).
- Gast, P. W. *Geochim. cosmochim. Acta* **32**, 1057–1086 (1968).
- Schilling, J. G. *Phil. Trans. R. Soc. A* **268**, 663–706 (1971).
- Langmuir, C. H., Bender, J. F., Bence, A. E. & Hanson, G. N. *Earth planet. Sci. Lett.* **36**, 133–156 (1977).
- De Paolo, D. J. & Wasserburg, G. J. *Geophys. Res. Lett.* **3**, 743–746 (1976).
- O'Nions, R. K., Hamilton, P. J. & Evensen, N. M. *Earth planet. Sci. Lett.* **34**, 13 (1977).
- Allègre, C. J., Ben Othman, D., Polvé, M. & Richard, P. *Phys. Earth planet. Inter.* **19**, 293–306 (1979).
- Loubet, M. *J. Petrol.* (submitted).
- Conquére, F. *Contr. Miner. Petrol.* **33**, 32–61 (1971).
- Wilshire, H. G. & Shervais, J. W. *Physics and Chemistry of the Earth* Vol. 9, 257–272 (Pergamon 1973).
- Hofmann, A. W. & Hart, S. R. *Earth planet. Sci. Lett.* **38**, 44–62 (1978).

Reversion of a promoter deletion in yeast

Stewart Scherer*, Carl Mann & Ronald W. Davis

Department of Biochemistry, Stanford University School of Medicine, Stanford, California 94305, USA

Promoter function in yeast has been examined by obtaining revertants of a his3 promoter deletion in vivo. Events which provide a new promoter for the his3 gene include insertion of the transposable element Ty1, rearrangements of the plasmid vector, and chromosomal mutations. A role for dicentric chromosomes as a source of the plasmid rearrangements is discussed.

PROMOTER function was originally defined by mutations which reduced the fully induced level of expression of the lactose operon of *Escherichia coli*¹. Since then, various selective methods have been developed to identify different types of altered promoter sequence. These include the isolation of mutations that increase gene expression and mutations which alter the regulation of specific genes². Additional information about promoter function in prokaryotes has come from studies of the *in vitro* interactions of purified RNA polymerase with promoter sequences³.

As most mutations that decrease gene expression, isolated *in vivo*, reside in coding sequences, genetic analysis of promoter function in eukaryotes has centred on systems where one can select for overproduction or altered regulation. An alternative approach to the study of promoter function is through the reversion of existing promoter mutations. Two eukaryotic promoter mutations have been examined in detail by this method: the *his4-912* and *his4-917* mutations of yeast⁴. In both cases, the pattern of revertants obtained by selection for restoration of *HIS4* function was influenced by the fact that the original mutations were due to the insertion of transposable elements.

Recombinant DNA techniques have made it possible to produce a series of deletions or base changes that modify promoter sequences. These altered promoters can then be returned to an appropriate host in order to examine their function *in vivo*⁵⁻⁷.

Here we have studied the reversion of a promoter deletion of the yeast *HIS3* gene. The deletion was first constructed in *E. coli* and then introduced into yeast as a new chromosome: a plasmid containing a functional centromere, an *ARS* sequence, and additional selectable markers (see Figs 1, 3). The promoter mutant chosen, *his3-Δ4*, deletes to within a few nucleotides of the start point of *HIS3* transcription (41 ± 1 nucleotides from the start of translation) in wild-type strains and produces no detectable *HIS3* mRNA⁷. The deleted sequences are replaced with λ DNA (see Fig. 3) rightward from the *att* site to the *Bam*HI site at 0.577.

Revertants of this *his3* promoter deletion are extremely rare, less than 1 in 10^9 viable cells, but arise by a variety of mechanisms. Yeast cells require only low levels of *HIS3* transcription for growth; this provides for a wide range in the level of expression that can be detected in revertants. In this report we describe several types of revertants. Potential sources for certain classes of the revertants and mechanisms for the restoration of *HIS3* function are discussed.

Blocking recombination between a plasmid-encoded gene and a chromosomal copy

Previous attempts to examine the reversion of *his3* mutant genes introduced into yeast by DNA transformation have been hampered by recombination with the *his3* allele present at the normal location on chromosome 15 (ref. 8). One solution to

this problem is to delete the homologous sequences at the normal *HIS3* locus. This approach has been used in the isolation of some of the revertants described below. In general, the removal of entire genes in yeast has been difficult. The interchromosomal regions are frequently small and suspected silent regions may encode essential functions for which transcripts are not readily detected.

Another solution to the recombination problem is to block all homologous recombination. However, in yeast, one does not have a single mutation, analogous to the *recA* mutants of *E. coli*, which effectively blocks most homologous recombination. The yeast *rad52* mutation⁹ will block gene conversion but reciprocal recombination remains unaffected. This latter form of recombination was effectively blocked in the *rad52* strains used here by incorporation of a yeast centromere^{10,11} into the plasmid. Cells containing the dicentric chromosomes which arise from reciprocal recombination of a centromere-containing plasmid into a normal yeast chromosome are not recovered in these strains.

The experiments which led to the above conclusion are summarized in Table 1. In each experiment, the host strain contained a *his3* gene with a small internal deletion and the plasmid contained the *his3-Δ4* promoter deletion. In the situation where the host strain was *RAD52*⁺ and the plasmid *cen*⁻, both gene convertants and reciprocal recombinants were recovered. The *cen*⁻ plasmids were unstable and in some cases where gene conversion events occurred, the other plasmid marker (and presumably the entire plasmid) had been lost. Cells that had integrated the plasmid by reciprocal recombination are at a selective advantage during the growth of the culture required to measure the recombination frequency. One would normally expect a rather higher ratio of gene conversion to reciprocal recombination than was observed¹². When the *rad52* mutation was incorporated into strains containing the *cen*⁻ plasmids, only the reciprocal events were recovered (Table 1). In addition, a high background of apparent suppressors of *his3-Δ4* was observed when using *cen*⁻ plasmids. This may be related to the poor control of copy number of the *his3* gene with these plasmids.

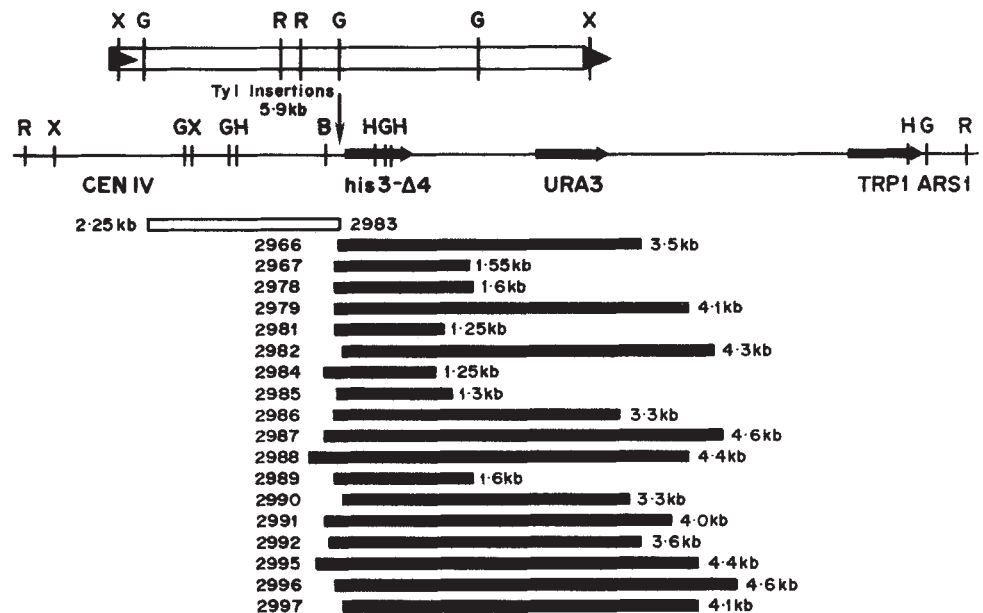
Table 1 Reversion to His⁺ of strains carrying the *his3* promoter deletion on a plasmid

Strain	Plasmid	Reversion frequency	Fraction Ura ⁺
<i>RAD52</i> ⁺	<i>cen</i> ⁻	$\sim 10^{-6}$	$\sim 40\%$
<i>RAD52</i> ⁺	<i>CEN</i> ⁺	$\sim 10^{-7}$	$\sim 50\%$
<i>rad52</i> ⁻	<i>cen</i> ⁻	$\sim 10^{-6}$	100%
<i>rad52</i> ⁻	<i>CEN</i> ⁺	$< 10^{-9}$	100%

The background of apparent suppressors, which generally show poor growth in selective conditions, has been excluded (see text). The *RAD52*⁺ strain was strain YNN36 (α *trp1-289 his3-Δ1 ura3-52*). The *rad52*⁻ strain was YNN209 (α *trp1-289 his3-Δ1 ura3-52 ade2 rad52*). The *cen*⁻ plasmid was 2944 (YRp12 with the *Bam*HI fragment containing *his3-Δ4*) and the *CEN*⁺ plasmid was 2961 (see Fig. 1) which contains a slightly smaller *his3-Δ4* fragment. Reversion to His⁺ of *his3-Δ1* strains containing no plasmid was $< 10^{-11}$.

* Present address: Division of Chemistry, California Institute of Technology, Pasadena, California 91125, USA.

Fig. 1 Plasmid rearrangements associated with reversion of the *his3* promoter deletion. A linearized map of plasmid 2961 with the positions of certain restriction endonuclease cleavage sites is shown. The total length of the plasmid is 11.1 kb. The ColE1 origin and ampicillin resistance gene are located between the *URA3* and *TRP1* genes. R, *EcoRI*; B, *Bam*HI; H, *Hind*III; G, *Bgl*II; and X, *Xho*I. Above the map of the plasmid is a map of the prototype *Ty1* element¹⁴. The terminal repeats of *Ty1*, the δ sequences, are indicated as filled triangles. One of the *Ty1* elements which inserted into 2961 lacked the *Xho*I sites in the δ sequences. The *Ty1* elements all inserted into the vector in the orientation shown. Transcription of *Ty1* is from right to left as drawn. The single open box represents the deletion contained in one of the revertants. The solid bars indicate the regions duplicated in the other revertants. The sizes of the deletion and the tandem duplications are shown. The sizes are ± 50 bp for the smaller rearrangements and ± 100 bp for the largest ones.



CEN⁺ plasmids in the *RAD52*⁺ strain will also convert *his3* mutations on chromosome 15 to His⁺. We did not determine whether plasmid integration by reciprocal recombination also occurred in this strain. Some of the His⁺ recombinants no longer contained the plasmid as judged by the loss of the other plasmid markers, which may be due to the imperfect stability of the centromere plasmids. It has also been noted that gene conversion events near centromeres are frequently associated with chromosome loss¹³. Finally, when the *CEN*⁺ plasmids were introduced into *rad52*⁻ strains, the frequency of His⁺ colonies fell dramatically. Among the His⁺ revertants, no His⁺ recombinants were detected (see below). In previous work¹², the block to gene conversion produced by *rad52* was not as complete as that observed here. However, in the experiments described here the regions of homology were quite small and one of the alleles contained an internal deletion. The frequency of viable recombinants in this situation is less than 1 in 10¹⁰ cells.

Revertants of the *his3*-Δ4 promoter deletion

All the revertants obtained in our study were derived using the same plasmid vector, 2961 (Fig. 1). Two host strains were used: a *rad52* strain, YNN209, which contains the *his3*-Δ1 allele⁸ and strain YNN211, which contains a total deletion of the *his3* gene, *his3*-Δ200 (M. Fasullo, unpublished strain). In both strains, the frequency of spontaneous His⁺ revertants was <1 in 10⁹ viable cells.

Spontaneous His⁺ revertants of yeast strains YNN209 and YNN211 containing plasmid 2961 fall into three broad classes: insertions of the transposable element *Ty1*, rearrangements of the plasmid DNA, and apparent chromosomal mutations. The revertants were first characterized by isolation of the plasmid DNA from the revertant cells by transformation of *E. coli*. Certain revertants yielded plasmids that appeared identical to the starting plasmid. All of these yeast strains grew poorly on selective media and the plasmids obtained from them would not transform yeast *his3* mutants to His⁺. These apparent chromosomal mutants have not been characterized further. The plasmid-encoded mutations are described below.

Reversion by insertion of the transposable element *Ty1*

Of 14 revertants isolated using the *rad52*/centromere-plasmid system described above, 6 were insertions of the transposable

element *Ty1*, as determined by comparison with restriction site maps of other cloned *Ty1* elements¹⁴ and confirmed by hybridization of ³²P-labelled *Ty1* sequences with the plasmid DNA recovered from the revertants (not shown). Restriction site mapping placed the insertion of the element near the 5' end of the *his3* coding sequences. The orientation of the element, as shown in Fig. 1, was the same relative to the gene as observed at the cytochrome *c*¹⁵ and alcohol dehydrogenase¹⁶ loci. The *Ty1* elements inserted into at least four distinguishable sites. This was determined by digestion of heteroduplexes of revertant DNA and end-labelled parental DNA with S₁ nuclease and measurement of the distance from the labelled end to the novel joint in the revertant DNA (Fig. 2a). These results are summarized in Fig. 3.

All six plasmids containing *Ty1* insertions transformed *his3* mutants to His⁺. The insertions which were closer to the start of the *HIS3* coding sequences expressed *HIS3* at a rather higher level as judged by growth in the presence of 10 mM 3-amino-1, 2, 4-triazole, an inhibitor of the *HIS3* gene product, imidazole glycerol phosphate dehydratase. Attempts to correlate the level of *HIS3* expression with insertion site were complicated by the fact that the *Ty1* elements that had inserted were not identical. Digestion with two restriction endonucleases, *Sal*I and *Xho*I, revealed that the presence of at least three different *Ty1* elements. Many polymorphisms of *Ty1* have been characterized^{14,17} and it is possible that the six insertions were six different *Ty1* elements. These studies have demonstrated that *Ty1* elements can transpose to a non-promoter sequence (λ DNA) near a gene and activate that gene. Also, *Ty1* transposition can occur in a *rad52* strain.

Of 21 revertants of *his3*-Δ4 obtained in a strain containing a total deletion of the chromosomal *his3* gene (YNN211); none were due to the insertion of *Ty1* elements. Thus, the frequency of transposition relative to the efficiency of other mechanisms available for reversion of the promoter deletion seems to be different for different strains. The source of the difference between these two strains is unknown.

Reversion by tandem duplication

Plasmids recovered from 15 revertants contained tandem duplications of a portion of the starting plasmid. Plasmids were screened for duplications by cleavage of the plasmid DNA with *Bgl*II endonuclease, which cleaves once in the *HIS3* gene and has no sites for ~6 kilobases (kb) downstream from this gene

(Fig. 1). As described below, all the duplications had one breakpoint near the 5' end of *HIS3* and one well beyond 3' of the gene in the vector sequences. These breakpoints produced a single novel junction near the 5' end of the *HIS3* gene for each new plasmid.

Cleavage of the revertant DNA with *Bgl*II produced a pattern of fragments that contained all the fragments of the vector with an additional fragment, which corresponds to the size of the duplicated region. Several examples of this are shown in Fig.

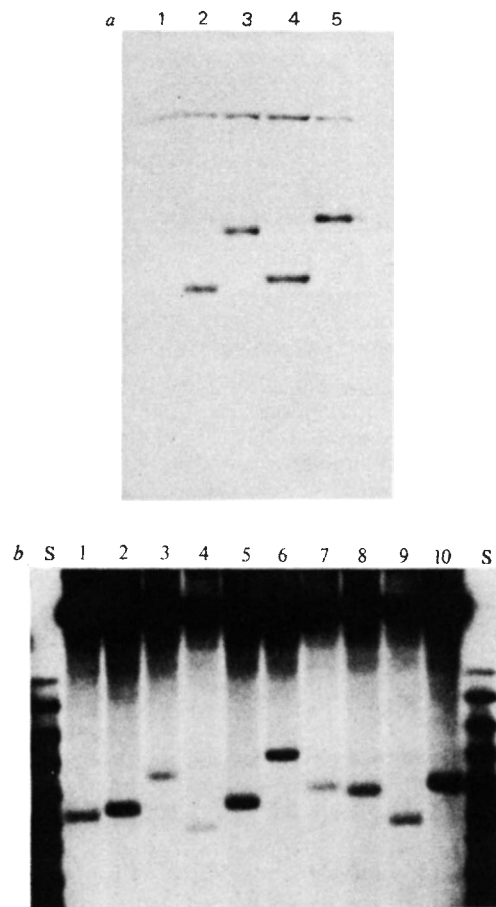


Fig. 2 Mapping of rearrangement breakpoints near *HIS3* with S_1 nuclease. 32 P-end-labelled DNA was prepared by treatment of *Hind*III-cleaved 2961 with calf intestine alkaline phosphatase followed by incubation with polynucleotide kinase and [γ - 32 P]-ATP. In *b*, the fragment of interest was purified by gel electrophoresis; in *a* it was cleaved also with *Bam*HI endonuclease before separation. For each sample, the DNA was cleaved with *Bgl*II endonuclease in 20 μ l of 10 mM Tris-HCl pH 7.5, 10 mM magnesium chloride, 1 mM dithiothreitol. In *a*, purified plasmid DNA was used; in *b*, DNA from rapid plasmid isolations was used. 3 μ l of 0.1 M Na_3 EDTA, 1 M NaOH and $\sim 10^4$ d.p.m. of probe fragment were added to each sample and incubated at room temperature for 15 min. The DNA was renatured by adding 3 μ l of 2 M Tris-HCl and incubating at 65 °C for 3–5 min. The samples were diluted with 300 μ l of 50 mM sodium acetate pH 4.4, 280 mM NaCl, 2 mM ZnAc₂ and digested with 1,000 units of S_1 nuclease at 37 °C for 30 min. After ethanol-precipitation, the samples were electrophoresed on a 5% polyacrylamide Tris-borate–7 M urea denaturing gel (*a*) or on a 2% agarose Tris-acetate–EDTA native gel (*b*). After electrophoresis, the gels were dried for autoradiography and exposed at –70 °C with an intensifying screen. The samples in *a* were: 1, no DNA; 2, 2970; 3, 2971; 4, 2973; and 5, 2976. The fragments were measured as 340–400 bp (standards not shown); the probe measures 575 bp. The samples in *b* were: 1, 2997; 2, 2996; 3, 2992; 4, 2990; 5, 2989; 6, 2987; 7, 2986; 8, 2985; 9, 2982; and 10, 2981. Standards, S, were *Bst*NI fragments of SV40 DNA.

4. Cleavage of the plasmid DNA from a revertant with any other enzyme which cleaves once in the duplicated segment will produce all the vector fragments normally observed for that enzyme and an additional fragment whose size is the same for all enzymes of this type.

The end point of the duplicated segment near the 5' end of the *HIS3* gene was localized by an S_1 nuclease mapping procedure similar to that used for the *Ty1* insertions (see Fig. 2*b*); the results are shown in Fig. 3. The rearrangement breakpoints were spread over a wide region beginning near the start of the wild-type *HIS3* mRNA and extending back as far as 500 base pairs (bp) from that point.

Given the size of the duplication and one of its end points, the other end point was determined. The regions that were duplicated in each revertant are shown in Fig. 1. There are no obvious hotspots for the duplication breakpoints, nor is there any obvious correlation between the duplications that break in a given region of the vector DNA and a particular distance from the 5' end of the *HIS3* gene for the other breakpoint. It is interesting that no duplications were found which fuse the *his3* gene to *URA3* sequences. The *URA3* gene is constitutively expressed and transcribed in the appropriate direction. About half of the revertants were selected in conditions where the *URA3* gene on the vector was also being selected. In these conditions, *URA3* is transcribed at a higher level¹⁸.

Eight of the duplication-containing plasmids, including those with the closest and the two most distant breakpoints relative to the *HIS3* gene, were tested for the ability of the plasmid DNA to transform *his3* mutant strains to His⁺. All transformants tested were His⁺. All except those generated using 2988 DNA were resistant to 10 mM 2-amino-1, 2, 4-triazole and therefore must express *HIS3* at near wild-type levels. The duplications were responsible for the *HIS3* expression as removal of the duplication from two different plasmids by homologous recombination resulted in reversion of the His⁺ phenotype to His[–].

Miscellaneous revertants

One revertant was a simple deletion of the starting plasmid; one breakpoint of the deletion was near the 5' end of the *HIS3* gene (Fig. 3). The size of the deletion was ~ 2.25 kb, which placed the other breakpoint in a region of yeast DNA normally transcribed into poly(A)-containing RNA¹¹. The deletion-containing plasmid retained at least partial *CEN4* function during mitotic growth.

Several revertants seemed to be duplications of most of the starting plasmid, excluding the *CEN4* region (not shown). The size of these duplications resulted in the production of a large number of coincidentally sized restriction fragments and made their characterization less certain.

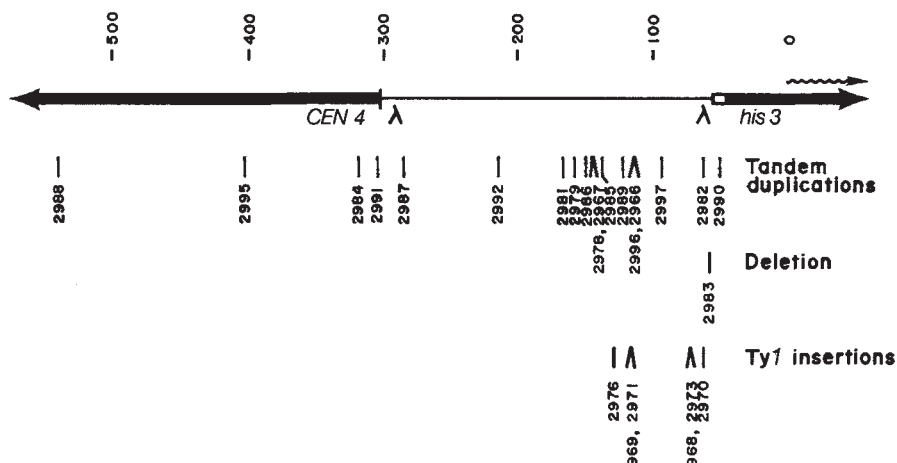
Two plasmids were found to contain multiple rearrangements of the starting plasmid DNA. One was a triplication or two similar duplications; the other appeared to contain two different duplications (not shown).

Plasmids isolated from the revertants described in this section transformed *his3* mutants to His⁺ and presumably represent the event in those cells that restored *HIS3* function.

Possible source of the plasmid rearrangements

Mutations associated with tandem duplications have been found in many systems. Some arise through recombination of repetitious DNA sequences and are generally sensitive to mutations affecting the general recombination systems of the organism. The large duplications of portions of the *Salmonella* chromosome are thought to arise by this type of mechanism¹⁹. In contrast, the frequency of spontaneous duplications of the phage λ chromosome is insensitive to such mutations²⁰.

Fig. 3 Rearrangement breakpoints near *HIS3*. The position of the breakpoint for the tandem duplications, deletion, and *Ty1* insertions (mapped as described in Fig. 2 legend) are shown. The positions are ± 10 bp for the closest breakpoints and ± 20 bp for the more distant ones. Distances from the start of the *HIS3* coding sequences (ATG) are given in base pairs at the top. The yeast DNA sequences in 2961 (*his3*- $\Delta 4$ and *CEN4*) are indicated as thick lines. The *CEN4* function is associated with sequences 2 kb to the left of the λ DNA (see Fig. 1). The *HIS3* structural gene sequences are indicated by the wavy line. The small region of λ DNA, to the right of the *att* site, attached to *his3*- $\Delta 4$ is shown as a thin line. The λ *att* site, which has partial homology to the yeast DNA that it replaced, is represented by an open box. All revertants shown were independently isolated, except 2966, 2967, 2968 and 2969, which form one group. The strains containing the rearranged plasmids are: pNN163 (2966), pNN164 (2967), pNN165 (2967), pNN166 (2969), pNN167 (2970), pNN168 (2971), pNN169 (2973), pNN170 (2976), pNN171 (2978), pNN172 (2979), pNN173 (2981), pNN174 (2982), pNN175 (2983), pNN176 (2984), pNN177 (2985), pNN178 (2986), pNN179 (2987), pNN180 (2988), pNN181 (2989), pNN182 (2990), pNN183 (2991), pNN184 (2992), pNN186 (2995), pNN187 (2996), pNN207 (2997). pNN163 to pNN172 were selected in YNN209, and pNN173 to pNN207 were selected in YNN211.



DNA rearrangements are normally quite rare in yeast. Some are due to recombination of repeated DNA sequences: rearrangements of chromosome 3 through the mating type loci²¹, recombination of *Ty1* sequences⁸ and recombination of the partially homologous cytochrome coding sequences²². Duplications have been detected in several cases. The *SAD1* mutation²³ duplicates a portion of the right arm of chromosome 3; it appears to be an unequal cross-over between *MAT* and *HMR*²⁴. Copper resistance is associated with tandemly repeated sequences (S. Fogel, personal communication). The results described here are unusual for yeast in that duplications seem to be the most abundant class of mutations.

We suggest that the formation of dimeric, dicentric plasmids may be an intermediate in the formation of the rearranged molecules found in the revertants described above. Asymmetric breakage of the dimer plasmids will be a source of deletions and tandem duplications (Fig. 5). The relative frequencies at which the revertants are recovered may reflect both the position of *his3* relative to the centromere in the starting plasmid and the locations of DNA sequences that can restore *HIS3* function.

With the isolation of centromeric DNA, it has become possible to construct dicentric chromosomes *in vitro*. In yeast, these

molecules are rapidly rearranged, usually by deletion of one of the centromeres (C. M., unpublished results). Plasmids that do not contain centromeres will freely form multimeric molecules in yeast²⁵; however, multimers of centromere plasmids are not readily detected¹¹.

Cyclization of the broken dimer plasmids leads directly to circular molecules containing simple duplications. Transformation of yeast with linear DNA molecules prepared from cloned

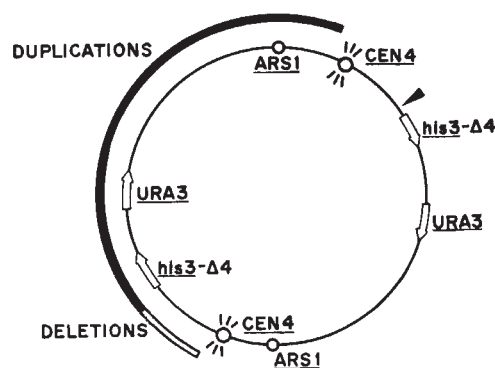


Fig. 5 A model for the origin of the duplications. A dimer of plasmid 2961 is shown. The various structural features of the plasmid are indicated. An arbitrary breakpoint near *his3* is marked by an arrowhead. The regions of the plasmid where a second break would yield a duplication (solid line) or a deletion (double line) are shown outside the plasmid.

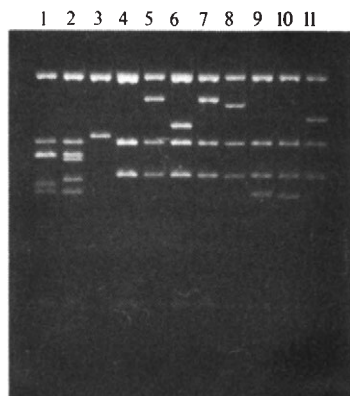


Fig. 4 Plasmid DNA from revertants. The samples were cleaved with *Bgl*II endonuclease and electrophoresed on a 0.8% agarose Tris-borate-EDTA gel. The samples in lanes 1 and 2 are 2976 and 2970, both *Ty1* insertions; lane 3, 2983, a deletion; lane 4, 2961, the parent plasmid; lanes 5–11, 2995, 2990, 2988, 2979, 2978, 2967, 2966 (all duplications). The sizes of the fragments of 2961 (lane 4) are: 6.15, 2.55, 1.90 and 0.55 kb (going from top to bottom).

circular DNA molecules has shown that these ends can be rapidly joined *in vivo* and that they appear to stimulate recombinational events involving the transforming DNA²⁶. One would expect that the molecules containing linear duplications produced by the mechanism outlined above would rapidly convert to circular molecules and possibly catalyse more complicated rearrangements. Broken dicentric chromosomes in maize are known to precede multiple rearrangements that can be observed cytologically²⁷. Consistent with this model is the observation that transformation of yeast with a dimer of plasmid 2961 yields >100 times the frequency of *His*⁺ colonies as that for the monomer. We chose to isolate the *His*⁺ revertants with the *HIS3* gene on a plasmid to facilitate rapid recovery of the functional *HIS3* DNA sequences. A different distribution of revertants would be predicted from the above model if the promoterless *HIS3* gene was located in a linear chromosome.

Creation of promoters by genomic rearrangement

In almost all the *HIS3* revertants described above, a new promoter was created by a genomic rearrangement near the 5' end of the gene. As can be seen with the tandem duplications, the distance from the start of the *HIS3* protein to the rearrangement breakpoint can be quite large, up to 500 bp.

The *Ty1* insertions are clustered more tightly than the duplications; however, the events that have been detected may reflect both the target specificity of the transposon, and the spectrum of sites from which it can activate a gene.

Activation of cryptic genes, or enhanced expression of wild-type functions by insertion of transposable elements, is found in many systems: for example, in *E. coli*, the major mechanism for activation of the genes for the metabolism of β -glucosides is the insertion of *IS1* and *IS5* (ref. 28). Similarly, the *c-myc* gene, the cellular homologue of the oncogene contained in the transforming virus MC29, is overproduced after nearby integration of a retrovirus²⁹.

Clearly, the orientation of the *Ty1* insertions is significant as all the insertions found here and all the insertions seen at cytochrome *c* (selecting for overproduction) and at alcohol dehydrogenase (selecting for altered regulation) are the same relative to the gene. This orientation is the reverse of that of transcription of the *Ty1* element itself³⁰. One might imagine bidirectional positive control (promoter) sequences at one end of the element. Some portion of this information must reside outside the identical end repeat sequences of *Ty1*⁸.

The duplications, which represent changes only in the arrangement of the plasmid DNA sequences, are more difficult to interpret. Most of the rearrangements involve only bacterial DNA sequences (λ and pBR322). Presumably several sequences which will pass for all or part of a promoter are to be found

in the vector DNA; however, they do not function in their present positions. The one known functional promoter, *URA3*, has not been fused to *his3* in any of the duplications examined. This may indicate some specificity in the rearrangement breakpoints or an incompatibility in the regulation of *URA3* with *HIS3* expression.

Duplication is an important mechanism in the evolution of chromosomes. It allows for new regulation of genes without loss of either the original copy of the gene or the source of the new regulation. One well-characterized example of this process is found in the mammalian globin genes. A series of related structural genes has evolved with different regulation for the various members of the family. Interspersed among these are several non-functional copies³¹. The new regulation developed after duplication need not be at the level of transcription. In the case of the immunoglobulin heavy chain constant region genes it appears to reside in additional genomic rearrangements³².

Although the mechanism suggested above as the source of the duplications applies to circular chromosomes, duplications clearly occur in linear chromosomes. The ability to insert the promoterless *his3* derivative at almost any point in the yeast genome allows one to examine the mechanisms which may produce duplications or other rearrangements in the chromosomes.

The experiments described here indicate that genomic rearrangements can create new promoters at low but detectable frequencies. As studies on the various aspects of gene expression proceed it will become increasingly important to understand the mechanics of the chromosomes on which they reside.

We thank Robert Malone for providing strains containing the *rad52* mutation. S.S. and C.M. were supported by NIH training grant funds. This work was supported by ACS grant no. NP-268B and NIH grant no. 21891.

Received 9 March; accepted 15 June 1982.

1. Scaife, J. & Beckwith, J. R. *Cold Spring Harb. Symp. quant. Biol.* **31**, 403–408 (1966).
2. Rosenberg, M. & Court, D. A. *Rev. Genet.* **13**, 319–353 (1979).
3. Losick, R. & Chamberlain, M. (eds) *RNA Polymerase* (Cold Spring Harbor Laboratory, New York, 1976).
4. Roeder, G. S., Farabaugh, P. J., Chaleff, D. T. & Fink, G. R. *Science* **209**, 1375–1380 (1980).
5. Benoist, C. & Chambon, P. *Nature* **290**, 304–310 (1981).
6. Fromm, M. & Berg, P. *Molec. cell. Biol.* (in the press).
7. Struhl, K. *Proc. natn. Acad. Sci. U.S.A.* **78**, 4461–4465 (1981).
8. Scherer, S. & Davis, R. W. *Science* **209**, 1380–1384 (1980).
9. Malone, R. & Esposito, R. *Proc. natn. Acad. Sci. U.S.A.* **77**, 503–507 (1980).
10. Clarke, L. & Carbon, J. *Nature* **287**, 504–509 (1980).
11. Stinchcomb, D. T., Mann, C. & Davis, R. W. *J. molec. Biol.* **158**, 157–179 (1982).
12. Jackson, J. A. & Fink, G. R. *Nature* **292**, 306–310 (1981).
13. Campbell, D. A. & Fogel, S. *Genetics* **85**, 573–585 (1977).
14. Cameron, J. R., Loh, E. Y. & Davis, R. W. *Cell* **16**, 739–751 (1979).
15. Errede, B., Cardillo, T. S. & Sherman, F. *Cell* **22**, 427–436 (1980).
16. Williamson, V. M. & Young, E. T. *Cell* **23**, 605–614 (1981).
17. Kingsman, A. J., Glimlich, R. L., Clarke, L., Chinnault, A. C. & Carbon, J. *J. molec. Biol.* **145**, 619–632 (1981).
18. Bach, M. L., Lacroute, F. & Botstein, D. *Proc. natn. Acad. Sci. U.S.A.* **76**, 386–390 (1979).
19. Anderson, P. & Roth, J. *Proc. natn. Acad. Sci. U.S.A.* **78**, 3113–3117 (1981).
20. Emmons, S. W., MacCosham, V. & Baldwin, R. L. *J. molec. Biol.* **91**, 133–146 (1975).
21. Strathern, J. N., Newlon, C. S., Herskowitz, I. & Hicks, J. B. *Cell* **18**, 309–319 (1979).
22. Ernst, J., Stewart, J. N. & Sherman, F. *Proc. natn. Acad. Sci. U.S.A.* **78**, 6334–6338 (1981).
23. Kassir, Y. & Herskowitz, I. *Molec. gen. Genet.* **180**, 315–322 (1980).
24. Nasmyth, K. A., Tatchell, K., Hall, B. D., Astell, C. & Smith, M. *Nature* **289**, 244–250 (1981).
25. Stinchcomb, D. T., Struhl, K. & Davis, R. W. *Nature* **282**, 39–43 (1979).
26. Orr-Weaver, T., Szostak, J. W. & Rothstein, R. J. *Proc. natn. Acad. Sci. U.S.A.* **78**, 6354–6358 (1981).
27. McClintock, B. *Genetics* **26**, 234–282 (1941).
28. Reynolds, A. E., Felton, J. & Wright, A. *Nature* **293**, 625–629 (1981).
29. Hayward, W. S., Neel, B. G. & Astrin, S. M. *Nature* **290**, 475–480 (1981).
30. Elder, R. T., St John, T. P., Stinchcomb, D. T. & Davis, R. W. *Cold Spring Harb. Symp. quant. Biol.* **45**, 581–591 (1980).
31. Proudfoot, N. & Maniatis, T. *Cell* **21**, 537–545 (1980).
32. Sakano, H., Huppi, K., Heinrich, G. & Tonegawa, S. *Nature* **280**, 288–294 (1979).

Chemical evidence for the capsomeric structure of phage Q β

Hisao Takamatsu & Koujiro Iso

Department of Chemistry, College of General Education, University of Tokyo, Komaba, Meguro, Tokyo 153, Japan

Novel capsomeric complexes, pentamers and hexamers were detected as chemical entities in phage Q β . Both were composed of identical protein subunits and stabilized by intermolecular disulphide bonds. Their numbers per particle were about 12 for pentamers and about 20 for hexamers—consistent with theoretical expectation from the quasi-equivalent packing of 180 identical subunits in a coat protein shell.

SINCE Loeb and Zinder discovered RNA bacteriophages¹, several spherical RNA-containing coliphages have been isolated and classified into four groups according to serological and physicochemical properties^{2,3}. Among them, groups I (for example, R17, f2 and MS2) and III (phage Q β) have been

extensively studied and have provided important insight into genetics and molecular biology.

A virus has a clearly defined and highly organized structure outside a cell, in contrast to the complexity of its form inside a cell. It was theoretically proposed that the protein

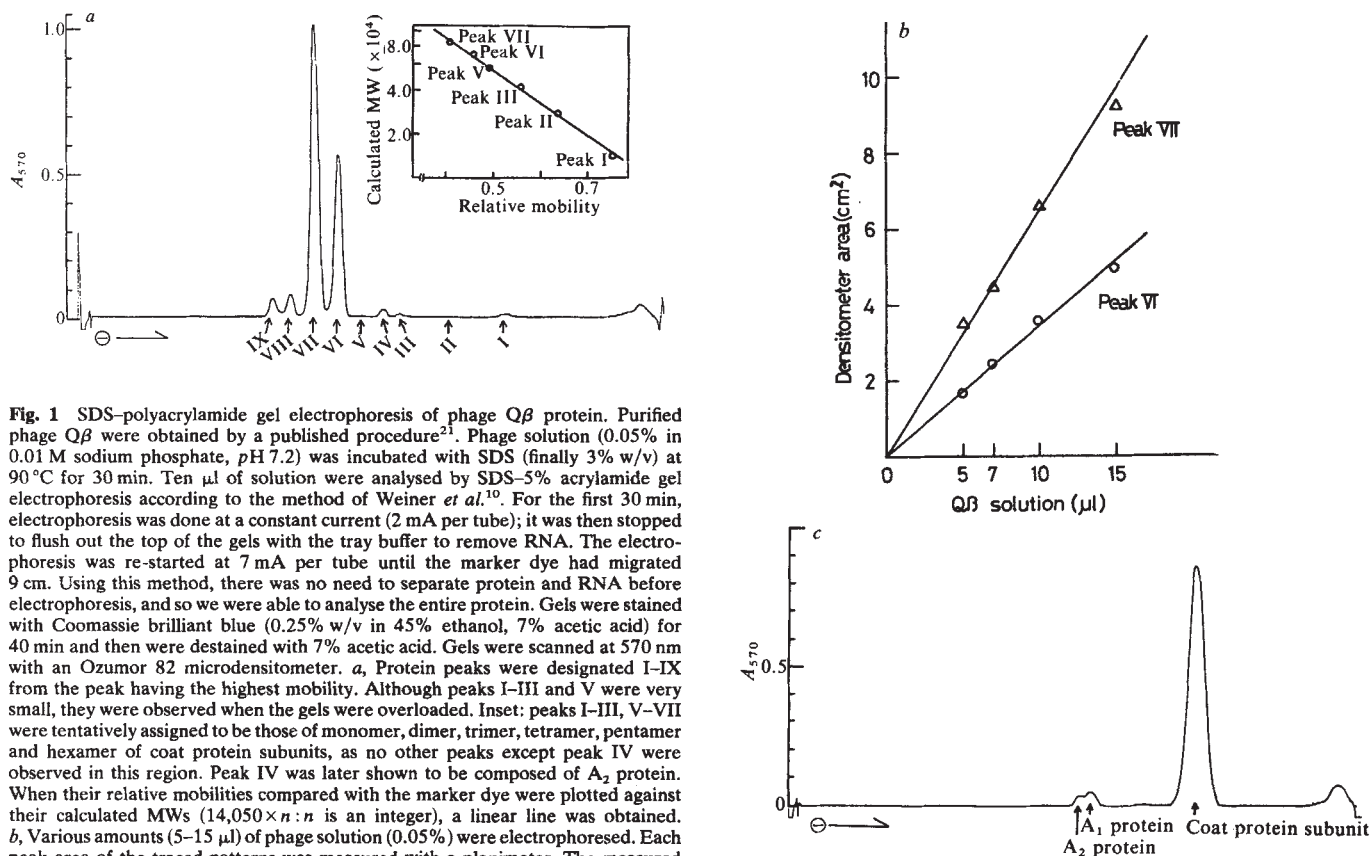


Fig. 1 SDS-polyacrylamide gel electrophoresis of phage Q β protein. Purified phage Q β were obtained by a published procedure²¹. Phage solution (0.05% in 0.01 M sodium phosphate, pH 7.2) was incubated with SDS (finally 3% w/v) at 90 °C for 30 min. Ten μ l of solution were analysed by SDS-5% acrylamide gel electrophoresis according to the method of Weiner *et al.*¹⁰. For the first 30 min, electrophoresis was done at a constant current (2 mA per tube); it was then stopped to flush out the top of the gels with the tray buffer to remove RNA. The electrophoresis was re-started at 7 mA per tube until the marker dye had migrated 9 cm. Using this method, there was no need to separate protein and RNA before electrophoresis, and so we were able to analyse the entire protein. Gels were stained with Coomassie brilliant blue (0.25% w/v in 45% ethanol, 7% acetic acid) for 40 min and then were destained with 7% acetic acid. Gels were scanned at 570 nm with an Ozumori 82 microdensitometer. *a*, Protein peaks were designated I-IX from the peak having the highest mobility. Although peaks I-III and V were very small, they were observed when the gels were overloaded. Inset: peaks I-III, V-VII were tentatively assigned to be those of monomer, dimer, trimer, tetramer, pentamer and hexamer of coat protein subunits, as no other peaks except peak IV were observed in this region. Peak IV was later shown to be composed of A₂ protein. When their relative mobilities compared with the marker dye were plotted against their calculated MWs ($14,050 \times n$; n is an integer), a linear line was obtained. *b*, Various amounts (5–15 μ l) of phage solution (0.05%) were electrophoresed. Each peak area of the traced patterns was measured with a planimeter. The measured areas were plotted against the applied quantities. Linearity was observed and the slope for peak VII was twice as steep as the slope for peak VI. Provided both peaks were composed of the same proteins, the weight ratio of peak VII to peak VI was 2.0. The same value was obtained after the gels were stained with Fast green which is also recommended for quantitative densitometry²². *c*, Phage solution was incubated with 2-ME (3% v/v) in the presence of SDS for 5 min at 90 °C. The reaction mixture was electrophoresed. The apparent MW was determined based on mobility relative to the marker proteins: ovalbumin (45,000), bovine serum albumin (67,000), chymotrypsinogen A (25,000) and myoglobin (17,800).

arrangement on the outer shell (capsid) of a virus could be predicted from a knowledge of the number of its constituent protein molecules⁴. For example, group I phage particles were estimated to contain about 180 molecules of coat protein, 1 molecule of maturation protein and 1 single-stranded RNA molecule⁵. The capsid structures were therefore considered to have a $T=3$ icosahedral symmetry⁴, and, indeed, electron microscopic analysis has shown that the coat protein subunits of phages R17 and f2 are in dimer clusters based on the $T=3$ structure⁶. Phage Q β is also predicted to have a $T=3$ icosahedral structure, as its protein composition is similar to that of group I phages, except for the presence of a few molecules of a third protein component⁷ (discussed below).

Here we have approached the determination of the protein arrangement (including minor protein components) in phage Q β , in two ways. First, by using a biochemical method, we have sought to characterize the constitutive intermediates during the assembly of protein subunits into the capsid. After the controlled dissociation of Q β particle, we have found pentamer and hexamer complexes, presumably true constituents of a $T=3$ capsid, which are stabilized through intermolecular disulphide linkages. The capsid is estimated to consist of ~ 12 pentamers and ~ 20 hexamers. Second, using electron microscopy, we show that the Q β capsid is composed of 32 morphological units (capsomeres) which we interpret to consist of 12 pentamers and 20 hexamers based on the quasi-equivalence theory⁴. The two approaches give consistent results.

This is the first report to describe the detection of a discrete substructure of a capsid, corresponding to a capsomere, as a chemical entity in a small spherical virus, and the verification of the quasi-equivalence theory by a chemical method. Detection of a substructure of a capsid may provide an important clue to the assembly process of a viral particle.

Specific dissociation of the phage Q β particle

The capsid is known to be stabilized in part by hydrophobic forces between coat proteins⁸, and the amino acid composition of coat protein is rich in hydrophobic residues⁹. We therefore reasoned that detergents or denaturing reagents might be useful for the disassembly of the capsid and the dissolution of the resulting products.

Phage Q β particles were treated with SDS and the reaction mixture was subjected immediately to SDS gel electrophoresis¹⁰. Nine protein peaks (I-IX, in order of decreasing mobility) could be observed reproducibly (Fig. 1a). About 90% of the total area is comprised of peaks VI and VII, which are estimated by comparison with several marker proteins to have apparent molecular weights (MWs) of 65,000 and 78,000 respectively. As a linear relationship exists between the applied quantity of protein and peak area (Fig. 1b), we can estimate the weight ratio of peak VI to peak VII; the ratio is 0.5. As the phage capsid is mainly composed of coat protein subunits⁵, the major peaks VI and VII probably consist of complexes of coat protein subunits.

Peaks I-III and V-VII are tentatively proposed to be monomer, dimer, trimer, tetramer, pentamer and hexamer of coat protein subunits respectively (peak IV is composed of A₂ protein, as described below). When the mobility of each peak is plotted against its calculated MW using the chemically determined MW (14,050)¹¹, a straight line is obtained (Fig. 1a inset). The pentamer of coat protein subunits (peak VI) and the hexamer (peak VII) are stable in SDS solution, so specific bonding must exist to keep these complexes intact; we suggest that it is disulphide bonding between protein subunits.

In addition to peaks VI and VII, there are seven small peaks in Fig. 1a. Peaks VIII and IX are particularly prominent and

together comprise ~8% of the total area. Their apparent MWs are 98,000 and 110,000 respectively. The other peaks amount to only ~2% of the total area. They can be distinguished when excess amounts of phage protein are separated by electrophoresis and their apparent MWs are 52,000 (peak V), 44,000 (peak IV), 39,000 (peak III), 28,000 (peak II) and 17,000 (peak I).

Reduction of the coat protein complexes

To confirm the disulphide bonding, phage Q β was treated with SDS in the presence of 2-mercaptoethanol (2-ME). As shown in Fig. 1c, three peaks can be observed, with mobilities corresponding to those of peaks I, III and IV. For convenience, the peaks have been designated peaks I, III and IV.

Peak I comprises ~94% of the total area, and peaks III and IV ~4% and ~2% respectively. The amount and relative mobility of each peak are consistent with published data^{7,12,13} and we conclude that peaks I, III and IV consist of coat protein subunit, A₁ (or IIB)⁷ protein and A₂ (IIA)⁷ protein respectively. The apparent MW of the coat protein subunit is estimated to be higher in our system than the chemically determined MW; similar values^{14,15} have been obtained by SDS-gel electrophoresis. These results show that our sample of phage Q β is of sufficient purity and further confirm that the components of low mobilities (as shown in Fig. 1a) are held by intermolecular disulphide bonds.

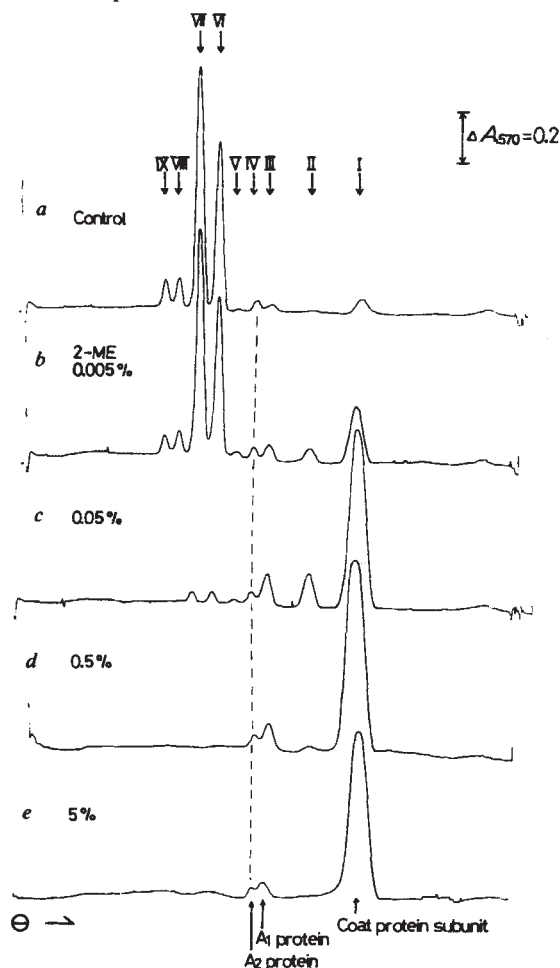


Fig. 2 Partial reduction of phage protein complexes. To the phage solution (0.05%, 10 μ l) was added an equal amount of 0%, 0.01%, 0.1%, 1.0%, or 10% (v/v) 2-ME solution and the mixture was incubated for 20 min at room temperature. Immediately afterwards, reaction mixtures were electrophoresed. After treatment with 0.005% 2-ME, peaks I–III and V were increased compared with the non-treated sample. When the pattern of the 0.05% 2-ME treated sample was compared with that treated with 0.005% 2-ME it was clear that peaks VI–IX had decreased and peak I had increased. In the presence of 0.05–5% 2-ME, peak I still increased, while peaks II, V–IX disappeared. Peak III decreased, but remained as in Fig. 1c. Peak IV was not affected by these treatments.

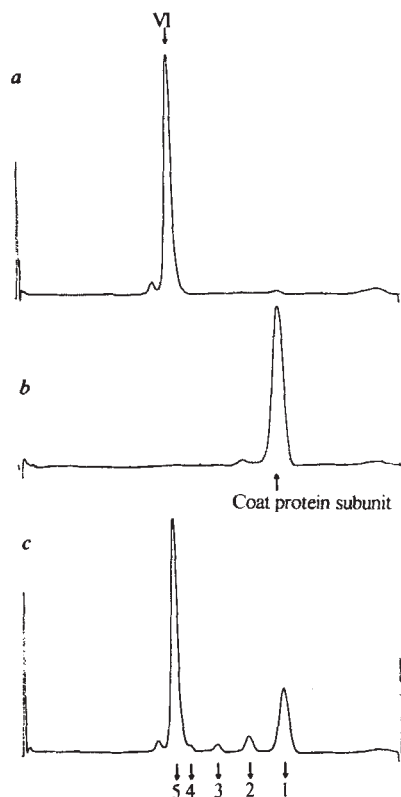


Fig. 3 Analysis of peak VI composition. Phage proteins were electrophoresed without treatment with 2-ME, as described in Fig. 1a legend. After electrophoresis, gels were frozen and sliced in 1-mm sections. Protein was then eluted from each slice into 0.2 ml of 0.1% SDS overnight at 37 °C with agitation. The protein solutions were divided into three parts: a, no further treatment and electrophoresed; b, treated with 5% 2-ME for 20 min at room temperature and then electrophoresed; c, treated with 0.005% 2-ME and immediately electrophoresed. As shown in a, the protein from peak VI was obtained. After treatment with 5% 2-ME, the peak of coat protein subunit was observed. In c, four new peaks can be seen, which have the same mobilities as peaks I–III and V in Fig. 1a. When the peaks were numbered as indicated by the arabic figures in c, peak VI was the fifth peak. No other peaks were detected in any other conditions. Peaks 1–5 had the same mobilities as peaks I–III, V and VI in Fig. 1a respectively.

We next tried to obtain an intermediate pattern between those of Fig. 1a and c. Phage Q β was treated with various concentrations of 2-ME in the presence of SDS. Significant changes could be seen (Fig. 2): further reduction decreased peaks VI and VII, and increased peak I. Peaks VIII and IX decreased in the same manner as peaks VI and VII. At the intermediate degree of reduction, peaks II, III and V increased slightly, in a reproducible fashion. These peaks were further diminished at greater concentrations of 2-ME. After complete reduction, only three constituent proteins of phage Q β could be observed. These results show that the complexes disassemble via intermediate products to the constituent protein subunits. These intermediates proved helpful in determining the structure of the complexes (see the following).

Structure of the specific complex

To determine the constitution of each peak, the gel was cut into slices and the protein eluted from each slice. The protein of peak VI was isolated, as shown in Fig. 3a. Reduction with 2-ME shifted the mobility of the protein peak to that corresponding to coat protein subunit (Fig. 3b). Thus peak VI was composed solely of coat protein subunits. The next question was how many coat protein subunits make up the complex of peak VI? As partial reduction of the complex led to the production of intermediates, in this case oligomers of coat protein subunits, it was possible to determine the polymerizing number of the original complex from the order of the intermediate oligomers. After mild reduction of the complex, four new peaks

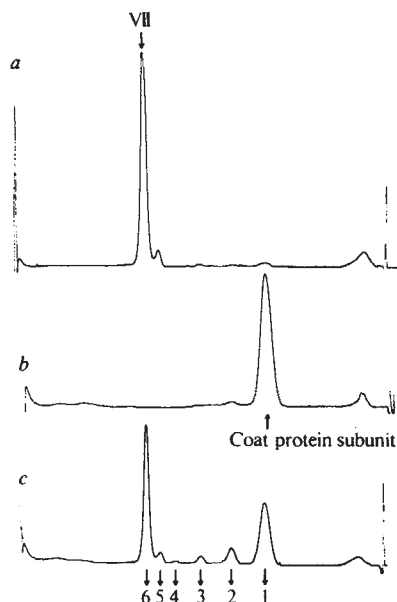


Fig. 4 Analysis of peak VII. The protein of peak VII was eluted from the sliced gel as described in Fig. 3 legend. The protein solution was treated with 0% (a), 5% (b) and 0.005% (c) 2-ME and then electrophoresed. After treatment with 5% 2-ME, only the peak corresponding to coat protein subunit was observed. Thus, peak VII was shown to consist of coat protein subunits. In c, six peaks were observed and no other peak could be detected in any other conditions. When the peaks were numbered as indicated by the arabic figures in c, peak VII was the sixth peak. Peaks 1–5 showed the same mobilities as corresponding peaks in Fig. 3c.

could be observed (Fig. 3c). No other peaks could be detected using different conditions, although greater reduction led to conversion of high molecular weight complexes to low molecular weight ones. Thus the four peaks are successive oligomers of coat proteins—monomer to tetramer. Therefore, the original complex (peak VI) must be a pentamer of coat protein subunits.

Peak VII (Fig. 4a) was analysed in the same way. Complete reduction leading to its full dissociation yielded only coat protein subunits, showing that peak VII consists of only coat protein subunits (Fig. 4b). With intermediate reduction, six peaks could be observed (Fig. 4c). No other peaks could be observed in different conditions, and the mobilities of the six peaks were the same as those of peaks I–III, V–VII in Fig. 1a. Thus peak VII consists of hexamers of coat protein subunits.

Next, peaks VIII and IX were examined. Because of their low mobilities, the separation distance between these peaks was relatively short compared with that between high-mobility peaks. Protein fractions eluted from the sliced gels therefore contained components of both peaks, in various ratios, depending on the position of the slices (peaks IX: VIII 10.1, 8.5, 3.0 and 0.7 in Fig. 5a–d). After treating each fraction with 2-ME, two peaks could be observed (Fig. 5a–d, lower patterns). The low-mobility peak corresponds to A_1 protein and the other to coat protein subunits. The molar ratios of coat protein subunit to A_1 protein are 4.6, 4.3, 4.1 and 4.9 in Fig. 5a–d respectively, assuming MWs of 39,000 for A_1 protein (electrophoretically determined) and 14,050 for coat protein subunit (ref. 11), and that the peak area is proportional to the applied quantity of protein, as shown in Fig. 1c. The high molar ratio for Fig. 5d is thought to be due to the contamination of peak VII (15%) which consists of coat protein subunits. On the basis of the molar ratios of A_1 protein to coat protein subunit, as well as the MWs of peaks VIII and IX, it is concluded that peak VIII consists of one A_1 protein molecule and four coat protein subunits, and peak IX consists of one A_1 protein molecule and five coat protein subunits. As the difference of the apparent MWs between peaks VIII and IX (12,000) equals that between peaks VI and VII (13,000), it is reasonable to conclude that the complex of peak IX is larger by one coat protein subunit than that of peak VIII.

Minor peaks I–III and V show the same mobilities as monomer, dimer, trimer and tetramer of coat protein subunits. It was observed, however, that the size of peak III increased with intermediate reduction and after full reduction it remained as an A_1 protein (Fig. 2). This suggested that peak III is a mixture of coat protein trimer and A_1 protein; this has been confirmed by a reduction experiment of the isolated components (not shown). Peak IV was not changed by treatment with various concentrations of 2-ME (Fig. 2), and after isolation, peak IV was shown to be composed of A_2 protein.

These results show that ~90% of the total phage protein is in the form of pentamers and hexamers, both of which are composed of only coat protein subunits. Accordingly, about 95% of all the coat protein subunits is calculated to be in the form of pentamers and hexamers since the fraction of the coat protein subunits in the whole phage protein is ~94%. The remaining coat protein subunits mostly form pentamers and hexamers with the A_1 protein molecules (peaks VIII and IX). Since the amount of A_1 protein is ~4% of the total phage protein and ~40% of peaks VIII and IX consists of A_1 protein, most of the A_1 protein molecules are in the form of pentamers and hexamers with coat protein subunits. A_2 protein does not form such a complex.

Capsid structure model

To formulate a model of the phage Q β capsid, we had to determine the contribution from each component. The total weight of protein per particle is 2.7×10^6 daltons, using the reported MW of 4.2×10^6 for the phage particle¹⁶ and 1.5×10^6 for its RNA¹⁷. About 90% of all the protein is in the form of pentamers and hexamers which are made up of the

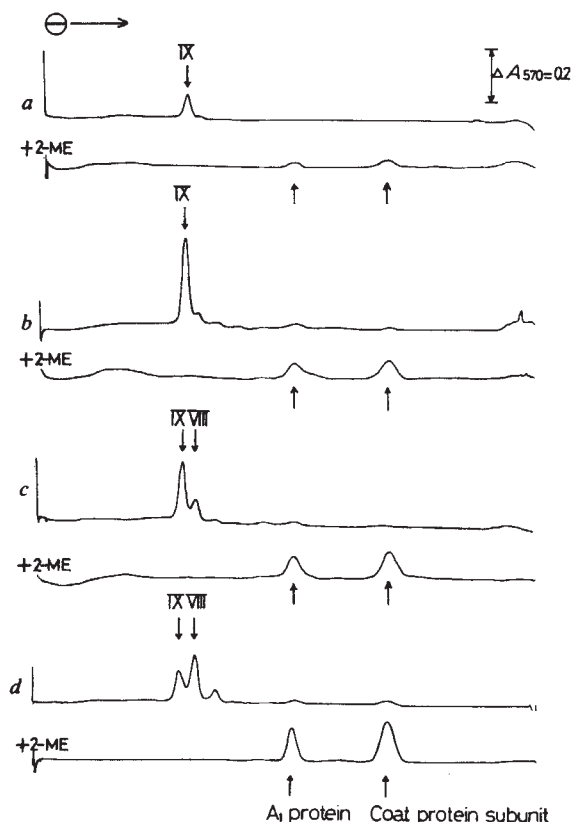


Fig. 5 Analysis of peaks VIII and IX. Protein was eluted from each sliced gel which contained the components of peaks VIII and IX. The ratio of components of peak IX to peak VIII were different in each slice: 10.1 (a), 8.5 (b), 3.0 (c) and 0.7 (d). There was a significant amount of peak VII (15%) (shown in d). The protein solutions were electrophoresed after treatment with 5% 2-ME (lower) or without treatment (upper). After treatment, two peaks were observed: the high-mobility one corresponds to coat protein subunit and the other to A_1 protein. The area ratios of coat protein subunit to A_1 protein were 1.64 (a), 1.54 (b), 1.49 (c) and 1.78 (d). The value for d was high due to the contamination of peak VII.

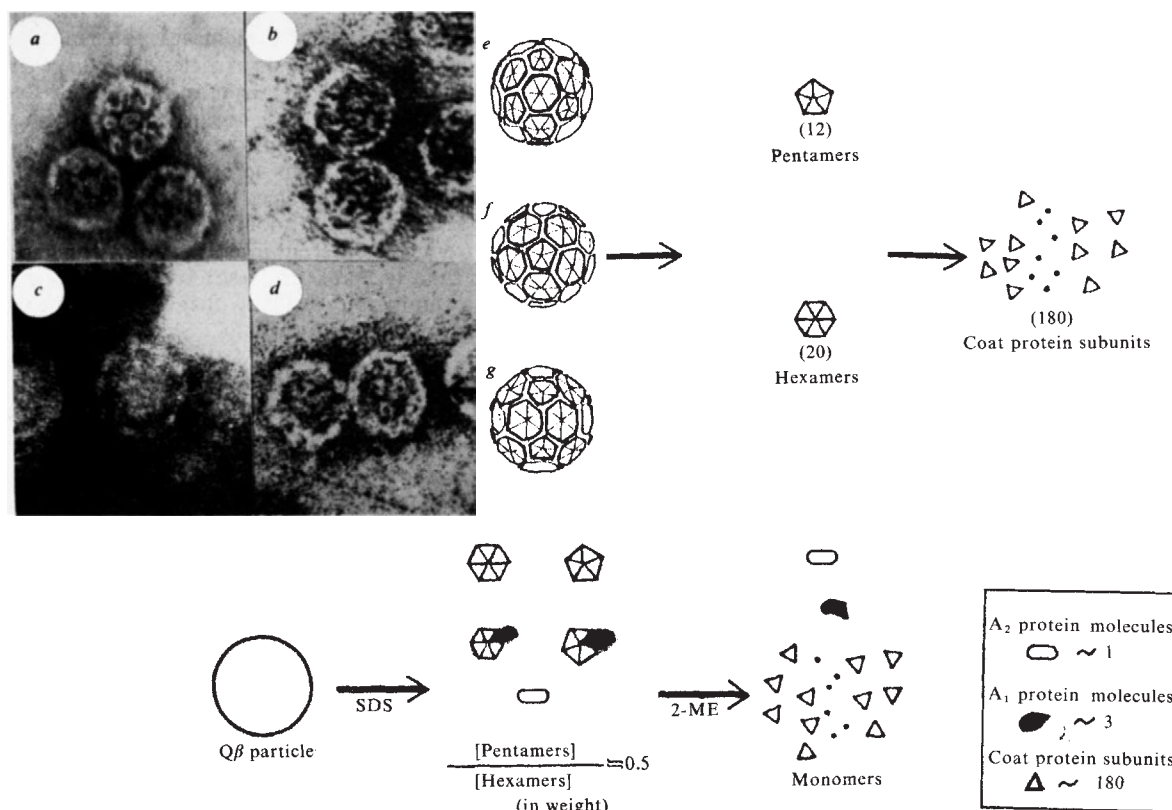


Fig. 6 Summary of electron microscopic (upper half) and chemical (lower half) observations. Phage Q β particles were observed by electron microscopy as described previously²³ although they were stained with phosphotungstate. The phage Q β was a spherical particle (mean diameter ~ 250 Å). On the particle surfaces, morphological units (capsomeres) were observed, which appeared as a hollow cylinder (~ 30 Å high and ~ 60 Å diameter with a ~ 25 Å diameter hole). Capsomeres were regularly arrayed on the particle in four ways: *a*, one capsomere surrounded by six capsomeres; *b*, one capsomere surrounded by five capsomeres; *c*, 10 capsomeres at the periphery of a particle; *d*, four capsomeres arranged at the four corners of a lozenge respectively. Based on the electron microscopy observations, the particle model was constructed, in which capsomeres were located at each apex (12) and on each facet (20) of an icosahedral structure. The capsomeres at 3-fold axes were drawn as hexagons and those at 5-fold axes as pentagons. The model is illustrated, viewed along the 3-fold axes (*e*), along the 5-fold axes (*f*) and along the 2-fold axes (*g*). According to the quasi-equivalence theory⁴, the capsomeres located at the 3-fold axes are hexamers and those at the 5-fold axes are pentamers. Then, the total number of subunits can be estimated to be $5 \times 12 + 6 \times 20 = 180$. Lower half: schematic drawings summarizing the chemical study. After phage Q β particles were treated with SDS, pentamers and hexamers of coat protein subunits were produced in a weight ratio of 0.5:1. In addition, A₁ protein-containing pentamers and hexamers were detected as minor products. These pentamers and hexamers were dissociated to subunits after reduction with 2-ME. It was calculated that one particle contained as protein components ~ 180 coat protein subunits, ~ 3 A₁ protein molecules and ~ 1 A₂ protein molecule.

coat protein subunits, and the weight ratio of pentamer to hexamer is 0.5. The number of these complexes per particle, therefore, is calculated to be 11.5 for the pentamers and 19.2 for the hexamers, based on the chemically determined MW of the coat protein subunit (14,050). On the other hand, A₁ protein-containing pentamers and hexamers comprise $\sim 8\%$ of the total amount of protein and are present in an equal amount per particle. There are thought to be about one of each per particle, based on their MWs ($95,000 = 39,000 + 14,000 \times 4$ for the pentamer; $109,000 = 39,000 + 14,000 \times 5$ for the hexamer). Therefore, it is estimated that, on the average, one particle contains ~ 12 pentamers and ~ 20 hexamers. In addition, the particle contains A₂ protein as a minor component ($\sim 2\%$ of total protein). As the MW of A₂ protein is 44,000, we estimate there is about one A₂ protein per particle; this is in good agreement with published data⁹.

The presence of 12 pentamers and 20 hexamers is consistent with a structure based on the $T = 3$ icosahedral surface lattice⁴, but the symmetry cannot be exact since only some of the pentamers and hexamers contain A₁ protein, and we do not know the location of the one copy of A₂ protein per particle.

Electron microscopy

The relationship between the chemically obtained complexes and the morphologically expected units on the capsid was investigated by observing Q β particles by electron microscopy. The Q β particle appeared as a regular, spherical particle

(~ 250 Å in diameter) with surface morphological units (capsomeres) consisting of hollow cylinders (~ 30 Å high and ~ 60 Å in diameter with a ~ 25 Å diameter hole) (Fig. 6). Capsomeres were regularly arrayed on the surface in one of four patterns: (1) one capsomere surrounded by the six nearest capsomeres (Fig. 6*a*); (2) one capsomere surrounded by the five nearest capsomeres (Fig. 6*b*); (3) ten capsomeres radially arranged at the periphery of a particle (Fig. 6*c*); (4) four capsomeres arranged at four corners of a lozenge respectively (Fig. 6*d*). These patterns would correspond to views of different sides of an icosahedral structure model as illustrated in Fig. 6, in which six capsomeres were arrayed around 3-fold axes (*e*), five capsomeres arrayed around 5-fold axes (*f*), 10 capsomeres located at the periphery (*f*) and four capsomeres arrayed in the 2-fold symmetrical position (*g*). From such consistency, it is concluded that the Q β capsid is composed of 32 capsomeres, arranged in an icosahedral symmetry (details will be published elsewhere by H. T., H. Funakoshi and K. I.). These morphological observations are consistent with the chemical estimation of the number of pentamers and hexamers per particle.

Protein subunits and icosahedral capsid structure

We have demonstrated that the pentamers and hexamers are stabilized by intermolecular disulphide linkage. As the coat protein subunit (131 residues) has two cysteine residues at positions 73 and 79 (ref. 11), pentamers and hexamers of coat

protein subunits can be held by successive disulphide bonds between subunits. Although we have no direct evidence of disulphide bonds in the particle, we believe this to be the case; even when Q β particles were disassembled at the low pH (pH < 4) or in the presence of an excess amount of iodoacetic acid¹⁸ most of the coat protein subunits were in the form of pentamers and hexamers, in the ratio of 0.5 pentamers to 1.0 hexamer. Furthermore, spectrophotometric studies using 5, 5'-dithiobis(2-nitrobenzoic acid) in the presence of guanidine hydrochloride or SDS¹⁹ showed most of SH groups were blocked (not shown). Pentamers and hexamers were clearly detected as chemical entities, although it remains to be determined whether the disulphide bonds are formed before or after virus assembly.

A₁ protein is known to be synthesized by the failure of termination of translation of coat protein cistron^{12,13}, but its

role is unknown, though the protein was reported to be essential for the formation of infectious particles²⁰. We have shown here that an A₁ protein molecule is substituted for a coat protein subunit in a pentamer and a hexamer. A₁ protein may bind coat protein subunits through the same contact between coat protein subunits since the N-terminus of A₁ protein molecule is identical to the entire stretch of coat protein subunit. A₁ protein may thus serve to modify the surface structure rather than to introduce a different intermolecular contact.

We particularly thank Professor Itaru Watanabe for giving us the opportunity to investigate phage Q β and for stimulating advice, also Drs K. Furuse, T. Shiba, H. Akanuma, H. Funakoshi, M. Moroi and S. M. Jung for helpful suggestions and Professor M. Yamasaki for encouragement. This study is supported in part by Scientific Grants from the Ministry of Education, Science and Culture of Japan.

Received 23 October 1981; accepted 25 May 1982.

- Loeb, T. & Zinder, N. D. *Proc. natn. Acad. Sci. U.S.A.* **47**, 282–289 (1961).
- Watanabe, I., Nishihara, T., Kaneko, H., Sakurai, T. & Osawa, S. *Proc. Jap. Acad.* **43**, 210–213 (1967).
- Miyake, T. *et al.*, *Proc. natn. Acad. Sci. U.S.A.* **68**, 2022–2024 (1971).
- Caspar, D. L. D. & Klug, A. *Cold Spring Harb. Symp. quant. Biol.* **27**, 1–24 (1962).
- Hohn, T. & Hohn, B. *Adv. Virus Res.* **16**, 43–98 (1970).
- Crowther, R. A., Amos, L. A. & Finch, J. T. *J. molec. Biol.* **98**, 631–635 (1975).
- Horiuchi, K., Webster, R. E. & Matsushashi, S. *Virology* **45**, 429–439 (1971).
- Matthews, K. C. & Cole, R. D. *J. molec. Biol.* **65**, 1–15 (1972).
- Weber, K. & Konigsberg, W. in *RNA Phages* (ed. Zinder, N. D.) 51–84 (Cold Spring Harbor Laboratory, New York, 1975).
- Weiner, A. M., Platt, T. & Weber, K. *J. biol. Chem.* **247**, 3242–3251 (1972).
- Konigsberg, W., Maita, T., Katze, J. & Weber, K. *Nature* **227**, 271–273 (1970).
- Moore, C. H., Farron, F., Bohnert, C. & Weissmann, C. *Nature new Biol.* **234**, 204–206 (1971).

- Weiner, A. M. & Weber, K. *Nature new Biol.* **234**, 206–209 (1971).
- Strauss, E. G. & Kaesberg, P. *Virology* **42**, 437–452 (1970).
- Furuse, K. *et al.* *Virology* **97**, 328–341 (1979).
- Overby, L. R., Barlow, G. H., Doi, R. H., Jacob, M. & Spiegelman, S. *J. Bact.* **91**, 442–448 (1966).
- Boedtker, H. & Gesteland, R. F. in *RNA Phages* (ed. Zinder, N. D.) 1–28 (Cold Spring Harbor Laboratory, New York, 1975).
- Glazer, A. N. in *The Proteins* Vol. 2 (eds Neurath, H. & Hill, R. L.) 1–103 (Academic, New York, 1976).
- Glazer, A. N., Delange, R. J. & Sigman, D. S. in *Chemical Modification of Proteins*, 113–114 (Elsevier, New York, 1975).
- Hofstetter, H., Monstein, H.-J. & Weissmann, C. *Biochim. biophys. Acta* **374**, 238–251 (1974).
- Watanabe, I., Miyake, T., Sakurai, T., Shiba, T. & Ohno, T. *Proc. Jap. Acad.* **43**, 204–209 (1967).
- Gorovsky, M. A., Carlson, K. & Rosenbaum, J. L. *Analyt. Biochem.* **35**, 359–370 (1970).
- Funakoshi, H. & Iso, K. *J. Biochem.* **70**, 201–204 (1971).

LETTERS TO NATURE

Is there a local source of magnetic monopoles?

S. Dimopoulos*, S. L. Glashow & E. M. Purcell

Lyman Laboratory of Physics, Harvard University, Cambridge, Massachusetts 02138, USA

F. Wilczek

Institute for Theoretical Physics, University of California, Santa Barbara, California 93106, USA

Cabrera¹ has reported the possible detection of a magnetic monopole in flight with magnetic charge g given by the Dirac condition² $2eg = hc$. Here, we accept the Cabrera candidate as a t'Hooft-Polyakov^{3,4} monopole of mass $M \sim 10^{16}$ GeV as expected⁵ in SU(5) or other grand unified theories. The monopole flux on Earth, on the basis of the single candidate, is $f \sim (0.1) \text{ cm}^{-2} \text{ yr}^{-1} (2\pi \text{ sr})^{-1}$, presumably consisting of roughly equal numbers of north and south monopoles. Galactic or intergalactic monopoles will have typical velocities of $\sim 300 \text{ km s}^{-1}$. The Cabrera flux would correspond to a mass density $fM/v \sim 1 \text{ GeV cm}^{-3}$. Because the mean mass density of the Universe cannot exceed $10^{-5} \text{ GeV cm}^{-3}$, the mean flux of monopoles in the Universe must be at least five orders of magnitude smaller than f (refs 6,7). This would not be a problem if the monopoles were concentrated in the Galaxy. However, Parker has shown⁸ that the existence of galactic magnetic fields is inconsistent with a mean galactic monopole flux of $10^{-7} \text{ cm}^{-2} \text{ yr}^{-1}$. It follows that f must be a local flux resulting from the special nature of the observation site. Three possibilities come to mind: the local monopole flux may be associated with the Earth, the Sun, or with the Solar System as a whole.

If the monopoles observed at the Earth's surface originate from the core, as does the local magnetic field, and if the flux

is not time dependent, the Earth has emitted more than 10^{27} monopoles since its birth, implying an initial (or present) monopole density $n \sim 1 \text{ cm}^{-3}$. However, the density of monopoles in the Earth's core is bounded by⁸

$$n < B/8\pi g v \tau \quad (1)$$

where $\tau = 10^4 \text{ yr}$ is the characteristic growth time of the geomagnetic field, v is the monopole velocity within the core, and B is the magnetic field within the core. With a core field $\sim 100 \text{ G}$, we expect $v \sim 3 \text{ km s}^{-1}$ on the basis of either magnetic or gravitational force arguments. This gives a limit $n < 10^{-9} \text{ cm}^{-3}$. Thus, the Cabrera flux cannot originate from the Earth.

Can the monopole flux seen on Earth originate from the Sun, in the fashion of radiant energy or solar wind? If so, the Sun has emitted (and presumably still contains) some 10^{36} monopoles. This implies a mean solar monopole density $n \sim 10^3 \text{ cm}^{-3}$, with 10^{-5} of the solar mass made up of monopoles. If the monopoles are gravitationally bound to the Sun, they must have velocities $\sim 400 \text{ km s}^{-1}$ and kinetic energies $\sim 10^{10} \text{ GeV}$. They will lose energy in nuclear collisions, and by inducing currents in the highly conductive solar medium. The energy they dissipate by these losses must be compensated by magnetic acceleration. Thus, there must be a mean interior magnetic field in the Sun of $\sim 10^3 \text{ G}$ so that

$$g\langle B \rangle \sim dE/dx \sim 10 \text{ MeV cm}^{-1} \quad (2)$$

The characteristic time of the solar interior field is assumed to be $\sim 20 \text{ yr}$, the period of the sunspot cycle. With these values of B and τ , equation (1) yields an upper bound on the density of monopoles in the Sun, $n < 10^{-7} \text{ cm}^{-3}$. Moreover, the ohmic loss due to 10^{36} orbiting monopoles is $\sim 10^5 L_\odot$. For both these reasons, it is difficult to explain the observed monopole flux in terms of a direct solar flux. Of course, it could be that most of the solar monopoles are concentrated in the centre of the Sun, and are only rarely extricated by transient magnetic fields. Escape velocity for these monopoles is $\sim 1,400 \text{ km s}^{-1}$, which would require magnetic fields $\sim 10^5 \text{ G}$ over distances $\sim R_\odot$.

Can the local monopole flux result from a diffuse cloud of monopoles which are in newtonian orbits about the Sun, rather

* Permanent address: Physics Department, Stanford University, Palo Alto, California 94305, USA.

like meteorites and meteoritic dust? Their velocities about the Sun will then be $\sim 30 \text{ km s}^{-1}$, while their arrival velocities at Earth will vary from 10 to 70 km s^{-1} as do the arrival velocities of meteorites. Their density must be $\sim 10^{-15} \text{ cm}^{-3}$ to explain the observed flux, corresponding to $\sim 10^{24}$ monopoles within the Earth's orbit. These monopoles are subject to a solar magnetic field $\sim 3 \times 10^{-5} \text{ G}$, and a consequent magnetic force which is $\sim 10^{-4}$ of the solar gravitational force. Following Parker⁹, we have estimated the time over which the magnetic perturbation significantly alters the energy of a monopole in solar orbit. We assume that the distant solar magnetic field fluctuates over a time scale of half a solar rotation period, $T = 10^6 \text{ s}$, and that its magnitude at 1 AU is $3 \times 10^{-5} \text{ G}$. The residence time of a monopole in solar orbit is given by⁹

$$t = (v/c)^2 T^{-1} \omega^{-2} \quad (3)$$

where $\omega = gB/Mc \sim 2 \times 10^{-15} \text{ Hz}$. We find $t \sim 100 \text{ Myr}$. The eventual fate of the orbiting monopoles will be to fall into the Sun, or to escape the Solar System. In any case, to maintain the local monopole flux, a source of $\sim 10^9$ monopoles per second is required.

We must verify that the monopole flux on the Earth does not generate too great a monopole density within the Earth. The incident monopoles have kinetic energy $\sim 10^8 \text{ GeV}$ and lose $\sim 1 \text{ MeV cm}^{-3}$ (hence $\leq 10^6 \text{ GeV}$) as they traverse Earth. They pass through the Earth without stopping. Their density, $n \sim 10^{-15} \text{ cm}^{-3}$, is well within the bound set by equation (1).

The monopole cloud of the Solar System is maintained either by an external (cosmic) or internal (solar) flux. Cosmic monopoles traversing the Sun may lose enough energy by ohmic losses or magnetic deceleration to be captured into solar orbit. Indeed, $\sim 10^9$ monopoles per second pass through a hypothetical sphere of solar radius if the Parker bound on the flux of galactic monopoles is saturated. The actual number impacting the Sun is larger by the factor $[(1 + (v_E/v)^2)]$ where v is the monopole velocity and $v_E \sim 600 \text{ km s}^{-1}$ is the escape velocity from the Sun. As $\langle v \rangle \sim 300 \text{ km s}^{-1}$, the enhancement is considerable, particularly for slower monopoles. The Sun may well capture enough monopoles to sustain its cloud.

On the other hand, the cloud may be fed internally by a flux of $\sim 10^9$ monopoles per second which is extracted from the Sun by solar flares. Either mechanism implies a solar monopole number of $\sim 10^{26}$ and a density of $n \sim 10^{-7} \text{ cm}^{-3}$, corresponding to an acceptable ohmic dissipation of $10^{-5} L_\odot$. With kilogauss mean fields, equation (2) is satisfied. The limit on monopole density given by equation (1) is just saturated, a result whose significance we discuss below.

Simple considerations of magnetic diffusivity yield the expression $T = (4\pi\sigma/c^2)R^2$ for the characteristic time of the solar magnetic field. This is known to overestimate $\bar{T}$ by six orders of magnitude⁸. In the presence of monopoles, $\bar{T}$ is correctly determined to be $\sim 20 \text{ yr}$ by the magnetic dissipation time given in equation (1).

Monopoles within the Sun can have other significant effects. They can lead to a lower central temperature by promoting heat transfer from the solar interior, or by catalysing nuclear fusion or nucleon decay. This could ameliorate the solar neutrino problem.

It is suggested that magnetic monopoles may have a substantial cross-section σ to induce nucleon-number violating processes such as



This process can be observed directly as monopoles pass through material on Earth; it can contribute to energy production in the Sun; and it can act as a source of high-energy solar antineutrinos which can be detected on Earth. Monopole-induced nucleon decay, seen in the laboratory, will not seem to conserve momentum because the recoiling monopole is not detected. A limit can be placed on σ from known limits on the nucleon lifetime: $f\sigma < \tau_p^{-1}$ where τ_p is the lower limit on the

proton lifetime in experiments where momentum conservation is not imposed. Taking $\tau_p > 10^{29} \text{ yr}$, we obtain $\sigma < 10^{-28} \text{ cm}^2$. With this cross-section, catalytic nucleon decay within the Sun can account for no more than $10^{-6} L_\odot$. However, this process can also yield GeV electron antineutrinos, with a flux on Earth as high as $10^3 \text{ cm}^{-2} \text{ s}^{-1}$. A search for this kind of solar emanation is called for.

The net flux of monopoles from all stars in the Galaxy would amount to 10^{28} monopoles per year. These will be accelerated by galactic magnetic fields to an escape energy of 10^{10} GeV , requiring an energy source of $10^{38} \text{ GeV yr}^{-1}$. This is easily compatible with the estimated energy content of the galactic magnetic field $(B^2/8\pi)V \approx 10^{57} \text{ GeV}$ and a restoration time of one galactic period, or 10^8 yr .

We conclude that the Cabrera flux of magnetic monopoles can be understood in terms of an orbiting cloud of monopoles in the Solar System, which is replenished by a small 'monopole wind' from the Sun or from the Galaxy. Monopoles may have an important role in solar energetics. Our hypothesis can be established if the monopole flux on Earth is shown to be similar (in velocity distribution, angular distribution, and its seasonal and diurnal variations) to the observed flux of meteorites, except for the fact that monopoles (but not meteorites) pass freely through the Earth.

We thank J. Preskill and R. Rosner for valuable discussions. This research is supported in part by the NSF under grant no. PHY77-22864.

Received 10 June; accepted 30 June 1982.

1. Cabrera, B. *Phys. Rev. Lett.* **48**, 1378–1381 (1982).
2. Dirac, P. A. M. *Proc. R. Soc. A* **133**, 60–72 (1931); *Phys. Rev.* **74**, 817–830 (1948).
3. 't Hooft, G. *Nucl. Phys. B* **79**, 276–284 (1974); **B105**, 538–547 (1976).
4. Polyakov, A. M. *Pis'ma Eksp. Teor. Fiz.* **20**, 430–433 (1974); *JETP Lett.* **20**, 194–195 (1974).
5. Georgi, H. & Glashow, S. L. *Phys. Rev. Lett.* **32**, 438–441 (1974).
6. Preskill, J. *Phys. Rev. Lett.* **43**, 1365 (1979).
7. Lazarides, G., Shafi, Q. & Walsh, T. F. *Phys. Lett.* **100B**, 21–24 (1981).
8. Parker, E. N. *Astrophys. J.* **160**, 383–404 (1970).
9. Parker, E. N. *Astrophys. J.* **139**, 951–958 (1964).
10. Wilczek, F. *Phys. Rev. Lett.* **48**, 1146–1149 (1982).
11. Callan, C. *Dyon-Fermion Dynamics* (Princeton University Preprint, May 1982).

Interstellar scintillation and pulsar velocities

A. G. Lyne & F. G. Smith

University of Manchester, Nuffield Radio Astronomy Laboratories, Jodrell Bank, Macclesfield, Cheshire SK11 9DL, UK

Interstellar scintillation of the intensity of pulsar radio signals is known to be the effect of diffraction in random irregularities in electron density, with typical scales of 10^{13} cm and density fluctuations of typically $3 \times 10^{-4} \text{ cm}^{-3}$. The observed fading pattern is believed to be due to the combined motion of the pulsar and the observer relative to an effectively stationary and quasi-static interstellar medium; this has, however, not been clearly demonstrated so far, and the possibility has remained open that the fluctuations are due at least partly to streaming or wave motions in the medium. Recent measurements of the proper motion of pulsars by radio interferometry¹ now allow a direct test. We present here new observations which show that the scintillation characteristics of pulsars are closely related to their velocities, as measured by their proper motions.

The basic theory of scintillation was presented by Scheuer², who analysed the effect of a thin screen of irregularities. An analytical solution for irregularities filling the whole path, and with a gaussian distribution of sizes, is due to Lerche³, who extended previous work by Lee and Jokipii⁴. (For a review of theory and observations see ref. 5.)

Table 1 Scintillation characteristics and velocities

PSR	Dis- persion measure (cm ⁻³ pc)	Dis- tance <i>D</i> (kpc)	τ_{408} (s)	B_{408} (MHz)	V_s (km s ⁻¹)	V_t (km s ⁻¹)
0301+19	15.7	0.56	335	0.23	47	100±11
0329+54	26.8	2.3	270	0.10	78	229±11
0355+54	57	1.6	110	0.034*	92	62±23
0525+21	50.8	2.0	100*	(0.14)*	230	141±180
0809+74	5.7	0.17	1,100	2.0†	23	41±5
0823+26	19.4	0.71	200	0.9†	174	365±10
0834+06	12.9	0.43	480	2.0	84	104±6
0950+08	3.0	0.09	1,700	(20)	34	15±3
1133+16	4.8	0.15	360	(7)	124	264±2
1237+25	9.3	0.33	660	4.0†	76	178±6
1508+55	19.6	0.73	100	0.06*	91	346±10
1604-00	10.7	0.36	1,260	0.9	20	12±17
1642-03	35.7	0.16	132	0.14*	49	36±11
1929+10	3.2	0.08	600	3.0†	36	33±2
1944+17	16.3	0.43	540	0.3	29	19±8
1952+29	20	0.20	780	0.9	24	41±10
2016+28	14.2	1.3	630	0.18	33	15±18
2020+28	24.6	1.3	360	0.5	98	97±18
2021+51	22.6	0.68	430	0.28	44	56±13
2217+47	43.5	1.5	150	0.13*	128	230±43

$V_s = 4.36 \times 10^4 (BD)^{1/2} \tau^{-1}$ V_t is the transverse velocity from ref. 1.

* Pulsars for which a separate series of observations was made at 930 MHz.

† Scintillation patterns exhibiting frequency drifting.

The velocity V of the pattern past the observer can be obtained from the characteristic time of fading τ provided that the scale S is known. Spaced receiver observations are of little use in determining S , as it is typically larger than the Earth's diameter. It is, however, closely related to the frequency structure of the pattern observed at a single site. This arises from the relation between multi-path propagation and θ , the angular spread of waves constituting the diffraction pattern. Lee and Jokipii⁴ show that the scale S is related to θ by

$$S = V\tau = 0.113 \lambda / \theta \quad (1)$$

and the correlation bandwidth Δf is related to θ by

$$\Delta f = \frac{c}{\pi D \theta^2} \quad (2)$$

where D is the distance of the pulsar. Hence at 408 MHz

$$V = 4.80 \times 10^4 (D \Delta f)^{1/2} \tau^{-1} \text{ km s}^{-1} \quad (3)$$

where the units of D , Δf and τ are kpc, MHz and s.

In equation (1) we have chosen the relation for a Kolmogorov power law distribution of irregularities, in contrast to a gaussian distribution: the coefficient differs only by $\sim 10\%$, but there is some evidence in favour of the power law even though there is as yet no complete analytical solution for diffraction in that case⁶. The bandwidth Δf in equation (2) is the frequency separation at which the correlation falls to 0.5.

We measured the characteristic fading time and correlation bandwidth at 408 MHz for 20 pulsars whose proper motions are known¹. Observations were made with the Mark 1A radio telescope, with an integration time of 1 min and frequency resolutions of 0.5 MHz and 0.125 MHz, using eight adjacent frequency bands. A typical observation on one pulsar lasted 3 hours. The bandwidth and time scales were found by fitting gaussian autocorrelation functions: this is formally correct only for the time scales as the autocorrelation function in frequency, although not precisely known, probably resembles the more complex function given by Lerche for the restricted case of a gaussian distribution of irregularities. We tabulate values of τ and of bandwidth B both as values where the autocorrelation falls to $1/e$ and assumed that $B \approx 1.2 \Delta f$.

In fitting gaussian functions we allowed both the height of the function and its zero level as degrees of freedom; this allows for broadband intrinsic variations in the pulsar signal both on short and long time scales, as discussed by Roberts and Ables⁶.

For some pulsars for which the frequency structure was not resolved with these bandwidths, a separate series of observations was made at 930 MHz using a frequency resolution of 0.5 MHz in eight adjacent bands; the time scales and bandwidths were scaled to 408 MHz assuming that $\tau \propto \nu$ and $B \propto \nu^4$. These pulsars are indicated in Table 1.

The scintillation pattern for some pulsars shows organized frequency drifting. For these, marked in Table 1, we have taken the bandwidth as the maximum, following the slope of the frequency drift. We find that values of τ and B generally repeat within $\sim 20\text{--}30\%$ from day to day. Some doubtful values are enclosed in parentheses in Table 1.

Figure 1 shows the relation between the observed transverse speed of the pulsar (from proper motion) and the quantity $4.36 \times 10^4 (BD)^{1/2} \tau^{-1}$ where D is the distance of the pulsar as used by Lyne *et al.*¹ in obtaining the proper motion. The theoretical relation is a line of slope unity through the origin.

Part of the correlation in Fig. 1 might arise artificially as a result of errors in the estimates of the distance D , as the ordinate contains a factor of D and the abscissa contains $D^{1/2}$. We therefore show in Fig. 2 the same results as an angular measure, that is as measured and expected proper motions. These quantities are entirely independent, and the correlation is seen to be at least as good as in Fig. 1.

The fact that there are roughly equal numbers of pulsars above and below the theoretical line in Fig. 2 is indicative of the success of the theory described earlier and of the model of the electron density distribution used to derive the distances⁷. The scatter in Fig. 2 is attributable to measurement errors of the proper motion, scintillation time scales and bandwidths as well as errors in the assumed distances. The small amount of scatter indicates that even in individual cases the distance estimates cannot be very much in error.

We note that there is a good relation in Fig. 1 between observed and calculated speeds. The lower speeds obviously can be affected by the Earth's orbital velocity of 30 km s^{-1} : we suggest that this accounts for the larger than expected scintillation speeds for those pulsars having proper motions less than about 30 km s^{-1} . In contrast we note that most of the derived

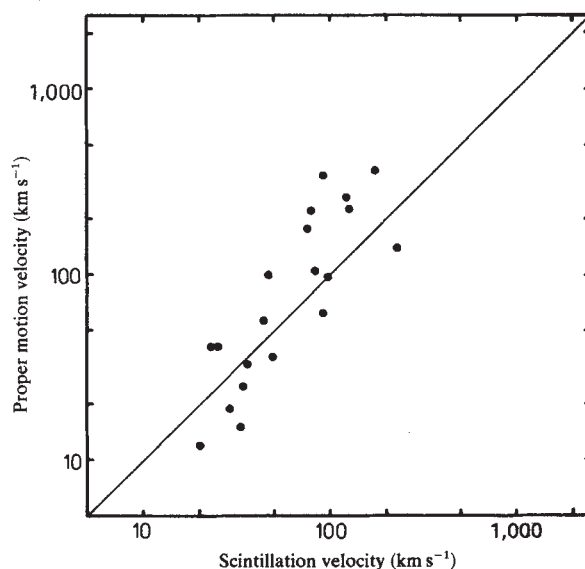


Fig. 1 The transverse speeds of 20 pulsars measured by their proper motions and compared with the speeds deduced from their scintillation characteristics.

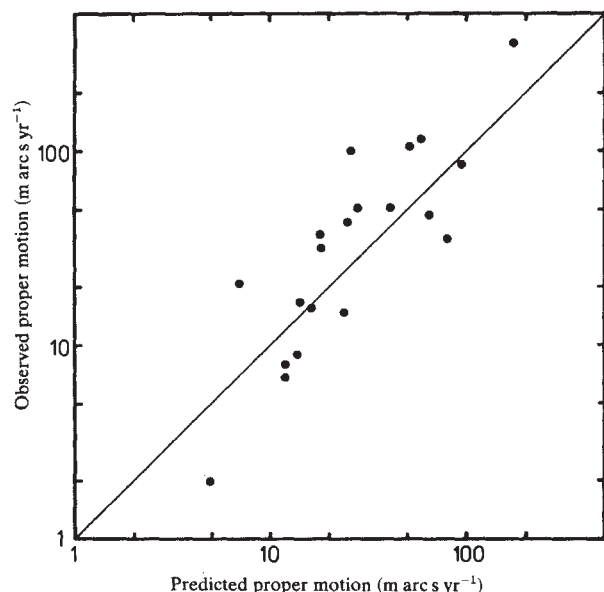


Fig. 2 The data of Fig. 1 expressed as angular motions: that is, a comparison between measured proper motions and angular speeds deduced from scintillation characteristics.

speeds of the faster moving pulsars are underestimates of the true values by a factor of about 2.

The agreement between theory and observation suggests that the irregularities are fairly evenly distributed between the pulsar and the observer. Most of the faster moving pulsars, for which the diffraction pattern is moving at about half the speed of the pulsar itself, are at large distances from the galactic plane: for these it is reasonable to suggest that the irregularities are concentrated within the half of the path closest to the observer. The main effect of such a concentration is a leverage which reduces the speed of the pattern by a factor of the order of 2.

These observations show that scintillation fading rates are directly related to the transverse velocities of the pulsars, with no appreciable contribution from velocities within the interstellar medium. This is in accordance with observations by Galt and Lyne⁸ who used simultaneous reception at widely spaced sites to obtain correct values for the magnitude and direction of the velocity of PSR0329+54. Some other observations, by Rickett and Lang⁹ and by Slee *et al.*¹⁰, also using spaced receivers, gave a contrary indication, and found apparent reversals of velocity in the diffraction pattern: however, as Hewish¹¹ pointed out, these may have been the effect of frequency drifts in the pattern combined with small differences in receiver frequency.

We now suggest that the apparent speeds deduced from scintillation measurements alone may provide a useful indication of the magnitudes of the actual speeds of the pulsars, even at distances so large that proper motion itself cannot be directly observed.

Inclinations of the uranian rings

R. G. French*†, J. L. Elliot*‡ & D. A. Allen§

* Department of Earth and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 01239, USA

† Department of Astronomy, Wellesley College, Wellesley, Massachusetts 02181, USA

‡ Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

§ Anglo-Australian Observatory, PO Box 296, Epping, New South Wales 2121, Australia

Rings formed with the birth of a planet are thought to arise after planetary condensation, from material remaining within the outer accretion limit—which has a radius of approximately two planetary radii, depending on the density of the material and the relative size of the particles^{1,2}. Although the initial distribution of particles would be spherical, collisions would cause rapid evolution of the particle orbits into circular orbits in the equatorial plane of the planet³. Also, gaps might form at the position of strong resonances with the planet's satellites, in an otherwise broad ring. Observations of planetary rings at high spatial resolution have revealed a more complex situation. The nine confirmed uranian rings are narrow, not broad, and most have elliptical orbits, not circular. Furthermore, the Voyager observations have shown that Saturn's rings consist of a multitude of narrow ringlets, at least some of which have elliptical orbits. Here we report an additional feature that we have found for the uranian ring orbits, not predicted by current theoretical models: at least seven of the nine rings have orbits inclined to the equatorial plane of Uranus.

The kinematic models for the uranian ring orbits have become increasingly sophisticated as more occultation timings have expanded the data base over the past 5 yr. The first kinematic model⁴ included only the α , β , γ and δ rings. It assumed their orbits to be circular, in a common plane that is coincident with the mean orbit plane of the five known uranian satellites. This would presumably be the equatorial plane of Uranus. The first improvement to the original model was made by Nicholson *et al.*⁵, who demonstrated that the ϵ ring is elliptical and precesses under the influence of the first spherical harmonic in the uranian potential field. Following this, elliptical models were formulated for all the known rings, and six of the nine were found to have measurable ellipticities⁶⁻⁹. The most complete model⁹ treats all nine rings as co-planar ellipses, precessing under the influence of the first two even harmonics of the uranian gravity potential; this model also includes the coordinates of the common orbit pole as free parameters. This model included 43 free parameters and was fitted by least squares to 105 occultation timings. The r.m.s. post-fit residual from this model is 0.31 s per degree of freedom, somewhat larger than the error of an occultation timing, which is <0.1 s in most cases. Hence, Elliot *et al.*⁹ concluded that this discrepancy implied that modelling the ring orbits as co-planar ellipses is not adequate and the situation must be more complex.

Although not expected on the basis of traditional models of ring formation³, the next step in a more complete kinematic model would include inclinations of the individual ring orbits to the equatorial plane of Uranus. For each of the nine rings we added two free parameters to our model: i , the inclination of the ring with respect to the uranian equatorial plane, and Ω_0 , the angle between the line of nodes of the ring orbit and a common reference line in the equatorial plane.

An inclined, elliptical ring undergoes both precession in its orbit plane and regression of its orbital nodes. The rate of

Received 7 June; accepted 1 July 1982.

1. Lyne, A. G., Anderson, B. & Salter, M. J. *Mon. Not. R. astr. Soc.* (in the press).
2. Scheuer, P. A. G. *Nature* **218**, 920 (1968).
3. Lerche, I. *Mon. Not. R. astr. Soc.* **189**, 137 (1979).
4. Lee, L. C. & Jokipii, J. R. *Astrophys. J.* **206**, 735 (1976).
5. Rickett, B. J. A. *Rev. Astr. Astrophys.* **15**, 479 (1977).
6. Roberts, J. A. & Ables, J. G. *Mon. Not. R. astr. Soc.* (in the press).
7. Manchester, R. N. & Taylor, J. H. *Astr. J.* **86**, 1953 (1981).
8. Galt, J. A. & Lyne, A. G. *Mon. Not. R. astr. Soc.* **158**, 281 (1972).
9. Rickett, B. J. & Lang, K. R. *Astrophys. J.* **185**, 945 (1973).
10. Slee, O. B. *et al. Mon. Not. R. astr. Soc.* **167**, 31 (1974).
11. Hewish, A. *Mon. Not. R. astr. Soc.* **192**, 799 (1980).

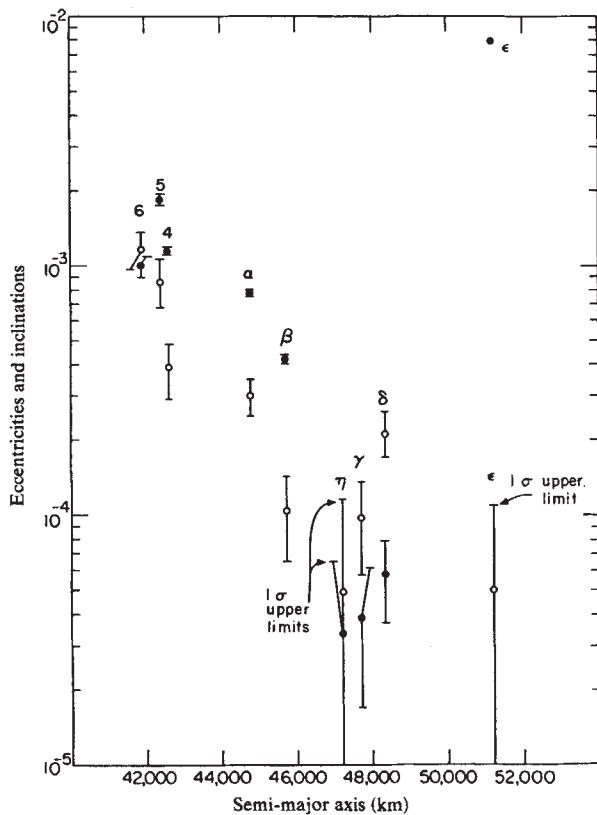


Fig. 1 Eccentricities (●) and inclinations (○) of the uranian rings. The eccentricity and inclination (in rad) for each of the nine presently known rings is plotted against the semi-major axis. Except for the eccentricity of the ϵ ring, the e s and i s show a decreasing trend with increasing semi-major axis. The eccentricities of the η and γ rings, as well as the inclinations of the η and ϵ rings, are not large enough to be statistically significant.

precession, $\dot{\pi}$, is given to second order in i , e and J_2 by the expression

$$\dot{\pi} = \left(\frac{GM}{a^3}\right)^{1/2} \left\{ \frac{3}{2} J_2 \left(\frac{R}{a}\right)^2 (1 + 2e^2 - 2i^2) - \frac{15}{4} J_4 \left(\frac{R}{a}\right)^4 \right\} \quad (1)$$

where the terms have their usual definitions⁹. Similarly, the rate of nodal regression, $\dot{\Omega}$, is

$$\dot{\Omega} = -\left(\frac{GM}{a^3}\right)^{1/2} \left\{ \frac{3}{2} J_2 \left(\frac{R}{a}\right)^2 (1 + 2e^2 - i^2/2) - \frac{1}{4} \left(\frac{R}{a}\right)^4 (15J_4 + 9J_2^2) \right\} \quad (2)$$

We fitted our expanded model to the available set of occultation timings, which included those used previously and a new set of 18 timings (see Table 1) from observation of the 26 April 1981 occultation by Uranus and its rings with the 3.9 m AAT¹⁰. The midtimes of the ring occultations were determined by methods used previously⁶. The results of the fit showed no substantial changes of the other parameters of the model⁹, but seven of the nine rings showed statistically significant inclinations—the largest being $(1.15 \pm 0.21) \times 10^{-3}$ rad for ring 6. The inclinations and eccentricities of the rings are plotted in Fig. 1. We note that the results for rings η and ϵ show no inclinations to the limit of our sensitivity. The r.m.s. post-fit residual per degree of freedom dropped from 0.31 to 0.14 s with inclinations included.

As a check on the consistency of our results, we also fitted the timing data with the precession and regression rates ($\dot{\pi}$ and $\dot{\Omega}$) of the elliptical and inclined rings as free parameters (not calculated from J_2 and J_4 as given by equations (1) and (2)). The values for $\dot{\pi}$ and $\dot{\Omega}$ obtained from this fit are plotted in

Fig. 2, and can be compared with the solid line plot of the rates calculated from J_2 and J_4 . The agreement is excellent.

Figure 1 shows that the inclinations generally decrease with increasing semi-major axis—a pattern noted earlier⁹ for the eccentricities, which remains in our present model. The ϵ ring has an exceptionally large eccentricity compared with the other rings, but its inclination is consistent with zero. For all rings (except ϵ) the value of $ae \sim ai$ (within a factor of 4). Also, our results show that $e > i$ for rings 5, 4, α , β and ϵ . For rings 6, γ and δ we find $i < e$, but this result lies within the uncertainty imposed by statistical errors.

The widest eccentric rings— α , β and ϵ —have been shown to be in uniform precession^{5,6}, probably by the mechanism of self-gravity¹¹. The uniform precession implies 'streamline' orbits, so that there is a gradient of as and es across the ring. Inclined rings in uniform precession would necessarily have inclination gradients within the ring. Such gradients would be more difficult to extract from the data than the eccentricity gradients, but we shall attempt to do so soon.

So far there has been little theoretical discussion about the significance of inclined rings. One way to form an inclined ring would be through the capture and tidal breakup of an incident, small body, which would almost certainly have an orbit inclined to the equatorial plane of Uranus^{12,13}. This explanation is highly unlikely for the inclined uranian rings, because the inclinations are so small. The shepherd satellite model for narrow rings¹⁴ can apparently account for orbital eccentricities¹⁵, and according to Goldreich and Tremaine (personal communication), orbital inclinations could be produced if the shepherd satellites have inclined orbits. However, the ability of this model to account for orbital inclinations has not been examined in detail.

Even after inclinations have been included in our orbit model, we note that the r.m.s. residual per degree of freedom (0.14 s)

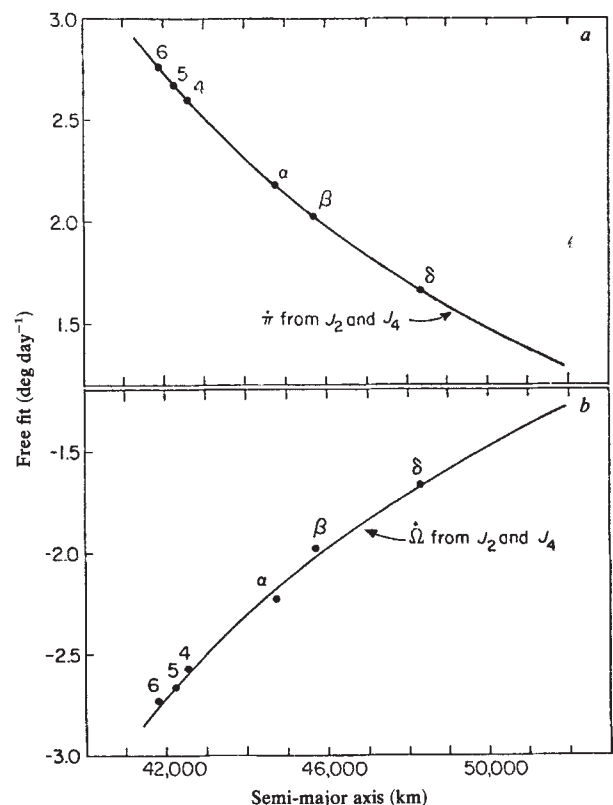


Fig. 2 Precession rates (a) and regression rates (b). The precession and regression rates for all rings having both a statistically significant inclination and eccentricity were fitted (not calculated from J_2 and J_4), simultaneously with all model parameters (except J_2 and J_4) allowed to vary in our previous fit (see Fig. 1). The results agree well with the rates calculated from equations (1) and (2), with fitted values for J_2 and J_4 (solid lines).

Table 1 Midtimes* of uranian ring occultations

Ring	Immersion	Emersion
6	19:18:56.56	20:17:50.43
5	19:18:31.03	20:18:11.02
4	19:18:19.24	20:18:27.74
α	19:16:30.51	20:20:18.76
β	19:15:42.38	20:21:05.74
$\eta^{\dagger}$	19:14:25.46	20:22:21.60
γ	19:14:02.63	20:22:44.16
δ	19:13:29.49	20:23:18.05
ϵ	19:11:10.02	20:25:28.10

* On 26 April 1981 UT.

 $\dagger$ The occultation times for the η ring refer to the narrow core at its inner edge⁹.

of our model is still significantly larger than the error in determining a ring occultation time. Hence, further improvements to our model should result in even more information about the Uranus system.

We thank A. Freedman, S. Tremaine and P. Goldreich for helpful discussions. This work was supported in part by NSF grant AST-8010699 and NASA grant NSG 7526.

Received 9 June; accepted 25 June 1982.

- Smoluchowski, R. *Nature* **280**, 377–378 (1979).
- Elliot, J. L. *et al. Nature* **294**, 526–529 (1981).
- Burns, J. A. in *The New Solar System*, 129–142 (Sky Publishing, Cambridge, Massachusetts, 1981).
- Elliot, J. L., Dunham, E., Wasserman, L. H., Millis, R. L. & Churms, J. *Astr. J.* **83**, 980–992 (1978).
- Nicholson, P. D., Persson, S. E., Matthews, K., Goldreich, P. & Neugebauer, G. *Astr. J.* **83**, 1240–1248 (1978).
- Elliot, J. L. *et al. Astr. J.* **86**, 127–134 (1981).
- Nicholson, P. D., Matthews, K. & Goldreich, P. *Astr. J.* **86**, 596–606 (1981).
- Nicholson, P. D., Matthews, K. & Goldreich, P. *Astr. J.* **87**, 433–447 (1982).
- Elliot, J. L. *et al. Astr. J.* **86**, 444–445 (1981).
- French, R. G. *et al. Icarus* (submitted).
- Goldreich, P. & Tremaine, S. *Astr. J.* **84**, 1638–1641 (1979).
- Smoluchowski, R. *Commun. Astrophys.* **8**, 69–78 (1979).
- Aggarwal, H. R. & Obserbeck, V. R. *Astrophys. J.* **191**, 577–588 (1974).
- Goldreich, P. & Tremaine, S. *Nature* **277**, 97–98 (1979).
- Goldreich, P. & Tremaine, S. *Astrophys. J.* **243**, 1062–1075 (1981).

An example of induced centrifugal force in general relativity

Jeffrey M. Cohen*, William J. Sarill†
& C. V. Vishveshwara‡

* Physics Department, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA

† Physics Department, Boston University, Boston, Massachusetts 02215, USA

‡ Raman Research Institute, Bangalore 560 080, India

Thirring's solution¹ for a rotating spherical mass is cited often as an example which exhibits induced Coriolis and centrifugal forces in general relativity^{2,3}. Previously, it was pointed out⁴ that the 'centrifugal terms' proportional to the square of the angular velocity actually represented quadrupole effects. These effects arose as a consequence of the latitude-dependent velocity distribution which generated an axially symmetric (non-spherical) mass distribution. This raises the question: Are there solutions to Einstein's equations which exhibit induced centrifugal force? If no such solutions exist, then the concept of induced rotation of inertial frames breaks down. Such a lack of connection between rotating masses and induced inertial frame rotation would in turn imply a breakdown of the

Machian^{5,6} idea that an inertial frame is a system of reference in which the stress-energy of the universe is at rest on the average⁷. Here we demonstrate that there are solutions to Einstein's equations which exhibit induced centrifugal force (other solutions may also be of interest in this connection⁸).

As pointed out elsewhere², velocity-dependent density distributions can give terms easily misinterpretable as centrifugal force—a difficulty which can be circumvented by treating a body for which the symmetry of the density distribution is velocity-independent; such a body is a rotating cylindrical shell. Other important advantages of this choice of source geometry are that it generates a simple geometry within the shell, that is, the flat space-time (ref. 9 and W.J.S., unpublished thesis), and the solution is exact. When written in the appropriate coordinates, flat space is well understood and easily interpreted. Furthermore, as the solution is exact, no ambiguities arise due to approximations.

A rotating cylindrical shell can be treated using a generalization of the Weyl-Levi-Civita^{10,11} metric

$$ds^2 = -e^{+2\psi} dt^2 + e^{2(\gamma-\psi)}(dr^2 + dz^2) + r^2 e^{-2\psi}(d\phi - \Omega dt)^2 \quad (1)$$

and the corresponding vacuum field equations

$$\begin{aligned} \nabla^2 \psi &= \frac{1}{2} r^2 e^{-4\psi} \nabla \Omega \cdot \nabla \Omega \\ \nabla \cdot (r^2 e^{-4\psi} \nabla \Omega) &= 0 \end{aligned} \quad (2)$$

derived by Ehlers¹² and later independently by the authors (ref. 13 and J.M.C. and W.J.S., unpublished). Here ∇ and ∇^2 are the usual flat space operators. The function γ is determined by quadrature from the equation

$$\gamma_r = r\psi_r - \frac{1}{4} r^3 e^{-4\psi} \Omega_r^2 \quad (3)$$

where the subscript denotes differentiation with respect to r and cylindrical symmetry has been assumed so that the derivatives with respect to z vanish. Inside the shell of uniform density, the only non-singular stationary exact solution is the flat space metric in rotating coordinates

$$ds^2 = -dt^2 + dr^2 + dz^2 + r^2(d\phi - \Omega dt)^2 \quad (4)$$

where Ω is a constant. This rotating form of the flat space-time is used to obtain an explicitly smooth match across the rotating shell; this form also facilitates an unambiguous identification of the Coriolis and centrifugal terms. The metric outside the shell can be obtained from the Weyl-Levi-Civita metric via the multivalued transformation $dt' = dt + K d\phi$ where K is a constant. The matching across the shell relates the constants Ω and K to the mass per unit length and angular velocity ω of the shell.

It is generally known that the geodesic equations corresponding to the flat space-time metric (3) contain Coriolis terms proportional to Ω and centrifugal terms proportional to Ω^2 . The frame in which all these terms vanish simultaneously is defined as the inertial frame. When the angular velocity ω of the shell vanishes, Ω also vanishes. Thus, Ω can be interpreted as the induced angular velocity associated with the dragging of inertial frames. We conclude, therefore, that there are solutions of Einstein's equations which exhibit induced centrifugal as well as Coriolis forces in accordance with Machian ideas.

This work was supported in part by the NSF.

Received May 28; accepted 18 June 1982.

- Thirring, H. *Phys. Z.* **19**, 33 (1918).
- Pauli, W. *Theory of Relativity* (Pergamon, New York, 1958).
- Einstein, A. *The Meaning of Relativity* (Princeton University Press, Princeton, 1950).
- Cohen, J. M. & Sarill, W. J. *Nature* **228**, 849 (1970).
- Brill, D. R. & Cohen, J. M. *Phys. Rev.* **143**, 1011 (1966).
- Sciama, D. W. *Mon. Not. R. astr. Soc.* **113**, 34 (1953).
- Dicke, R. H. *Scient. Am.* **47**, 25 (1959).
- Cruz, V. de la & Israel, W. *Phys. Rev.* **170**, 1187 (1968).
- Davies, H. & Caplan, T. A. *Proc. Camb. phil. Soc. math. phys. Sci.* **69**, 325 (1971).
- Weyl, H. *Rc. Accad. Lincei* **54**, 117 (1917).
- Levi-Civita, T. *Rc. Accad. Lincei* **26**, 381 (1917).
- Ehlers, J. thesis, Univ. Hamburg (1958).
- Vishveshwara, C. V. & Winicour, J. J. *Math. Phys.* **18**, 1280 (1977).

Space shuttle ice nuclei

R. P. Turco*, O. B. Toon†, R. C. Whitten†
& R. J. Cicerone‡

*R & D Associates, Marina del Rey, California 90291, USA

†NASA Ames Research Center, Moffett Field,
California 94035, USA

‡National Center for Atmospheric Research, Boulder,
Colorado 80307, USA

With the advent of the space shuttle era, increasing rocket activity is expected in the Earth's upper atmosphere. The space shuttle solid-fuelled booster rockets emit ~150 tonnes of alumina (Al_2O_3) particles per launch, much of which spreads throughout the upper troposphere and stratosphere¹. Such particles can seed aerosols and clouds, and might therefore cause changes in the terrestrial radiation balance and climate². Estimates are made here showing that average ice nuclei concentrations in the upper troposphere could increase by a factor of 2, and that an aluminium dust layer weighing up to 1,000 tonnes might eventually form in the lower stratosphere.

The morphology of alumina particles collected in high-altitude Titan exhaust trails may be summarized as follows^{1,3}: typically smooth and spherical with occasional deeply-grooved surface structures; hexagonal α - Al_2O_3 or cubic γ - Al_2O_3 crystallography; heavy-element composition dominated by Al with smaller amounts of K, Na, Ti, Fe and Si; trace quantities of chlorine adsorbed on surfaces; and densities in the range of ~ 1.5 to $\sim 3.5 \text{ g cm}^{-3}$, suggesting individual particles with a porous or hollow construction.

Previous theoretical investigations⁴⁻⁶ focused on the role of Al_2O_3 particles as stratospheric condensation nuclei (CN, considered here as particles capable of nucleating supersaturated sulphuric acid vapour). However, space shuttle Al_2O_3 particles emitted in the stratosphere eventually filter downwards into the troposphere, after being widely dispersed over the hemisphere of injection. Inasmuch as alumina provides an excellent ice-forming epitaxy⁷, they might contribute significantly to the population of background ice deposition nuclei (IN) at high altitude. IN produce ice crystals directly from the vapour phase at low temperatures (-5° to -25°C) and supersaturations ($\geq 10\%$, or equivalently, relative humidities over ice ~ 100 – 110%). Cloud condensation nuclei (CCN), which can nucleate liquid water droplets at supersaturations $\leq 1\%$, are not subject to great influence by space shuttle emissions on a global scale.

Parungo and Allee^{8,9} and Hindman and co-workers¹⁰⁻¹³ studied IN concentrations in the stabilized ground clouds of solid-fuelled Titan rockets, which emit copious quantities of Al_2O_3 dust. They also discuss related laboratory data. Recognizing that the conditions in ground clouds are quite different from the conditions in high-altitude launch plumes, a consensus of the ground-cloud research nevertheless points to copious production of IN by rocket engines, perhaps as many as 10^{10} IN per gramme of fuel burned (active at -16° to -20°C and 101% relative humidity over ice).

To estimate the accumulation of space shuttle IN (or SSIN) in the upper troposphere and lower stratosphere, a comprehensive one-dimensional model of multicomponent atmospheric aerosols was used^{14,15}. The model segregates sulphate aerosols and un-nucleated CN particles, and keeps track of the amounts of different CN materials incorporated into the aerosols by nucleation and coagulation. In this manner, the Al_2O_3 component of aerosols in the size range from 0.001 to $10 \mu\text{m}$ radius (subdivided into 40 distinct size bins) was calculated for the size- and height-dependent Al_2O_3 particle deposition rate associated with the space shuttle. The aerosol physics and chemistry treated in the model are discussed elsewhere¹⁴. In the present simulations, micrometeorites, which are the dominant natural stratospheric particles above $\sim 1 \mu\text{m}$ radius, are neglected; their presence does not materially alter the results.

The adopted Al_2O_3 mass deposition profile corresponds to 52 space shuttle launches per year, with emissions averaged over the Northern Hemisphere and smoothed in time to yield a continuous injection rate⁴. All of the aluminium burned was assumed to condense as dust. Particle interactions in the rocket wake were ignored, as these interactions act mainly to reduce the particle population below $\sim 0.01 \mu\text{m}$ radius (see below). Al_2O_3 particle injection above 8 km was included in the calculations, the total mass deposition amounting to $\sim 5,000$ tonnes yr^{-1} . The size distribution of the injected particles (the number injected per cm^3 of air per μm radius) was taken to be constant below $0.01 \mu\text{m}$ radius, and to vary as $r^{-\alpha}$ above $0.01 \mu\text{m}$. The latter dependence, with α in the range 3.5–4.5, is suggested by direct sampling in Titan exhaust plumes^{1,8,9}. The average particle density was assumed to be 1.7 g cm^{-3} , which emphasizes the number deposited.

The Al_2O_3 particle nucleation time in the stratosphere, τ_n , and the aerosol 'rainout and washout' lifetime in the troposphere, τ_r , were parameterized in a simple manner. For aluminium oxide dust, a fixed, relatively short nucleation time of 100 s was assumed as a base case¹⁶. Longer nucleation times up to 10^6 s were also tested. The simulated rainout loss rate decreases linearly from the ground-level value τ_r^{-1} to zero at the tropopause¹⁴. The baseline rainout rate of 10^{-6} s^{-1} at the ground translates into a rainout rate of $\sim 10^{-7} \text{ s}^{-1}$ near 10 km; ground-level rates as small as 10^{-7} s^{-1} were also considered.

To estimate the concentrations of SSIN in the upper troposphere and lower stratosphere using model predictions, three factors were taken into account: the composition of the particles, the extent of surface poisoning, and the size of the particles. The average composition of the sulphate aerosols in a given model size bin, i , is specified by the material volume fractions, $f_i(j)$, where j is a material index. Similarly, the volume fraction of material j in CN of size i is $f_i^{\text{CN}}(j)$. In the present work, the aerosol materials consist of three types: sulphuric acid aqueous solution, 'tropospheric' particulates, and aluminium oxide dust. The CN are composed of only the last

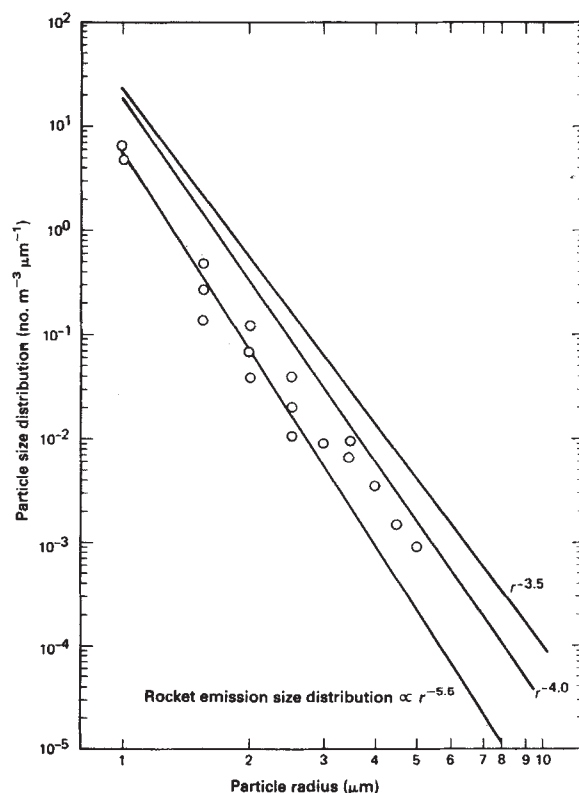


Fig. 1 Calculated stratospheric aerosol size distributions at 20 km with space shuttle particulate injection. 52 shuttle launches yr^{-1} , with emissions averaged over the Northern Hemisphere, were assumed. In each calculation, a nucleation time of 100 s was used (see text). Observations are compared: $\circ$, data from ref. 3 for Al_2O_3 at 20 km, multiplied by 10.

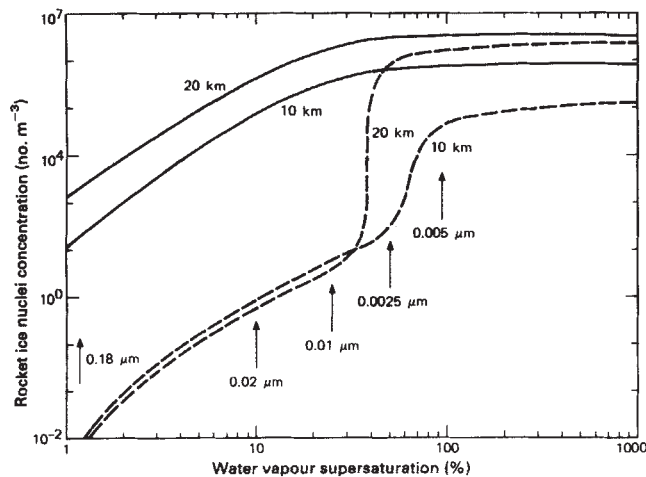


Fig. 2 Calculated space shuttle ice nuclei concentrations at 10 and 20 km altitude corresponding to different water vapour/ice supersaturations. The results were obtained for 52 shuttle launches per year with emissions averaged over the Northern Hemisphere, a dust emission size distribution of $r^{-4.5}$, and a nucleation time constant (τ_n) of 100 s. Estimates of the maximum (solid line) and average (dashed line) IN concentrations are given (see text). Also indicated are the minimum radii of spherical ice nuclei for several water vapour supersaturations at -30°C .

two materials. The maximum number of CN identified as active SSIN in each model size bin was taken as the fraction, $f_i^{\text{CN}}(\text{Al})$, of the total number of CN in that size bin. Nucleated CN are classified as acid aerosols in the model, and are discussed below. Possible deactivation of Al_2O_3 CN by coagulation with background 'tropospheric' CN was neglected in the stratosphere, as the background CN concentrations are so low ($<10\text{ cm}^{-3}$).

For H_2SO_4 aerosols, two estimates of the SSIN concentration were made (and added to the CN SSIN concentrations derived above). They are a 'maximum' number and an 'average' number. Such an approach was called for because the precise extent of the poisoning of individual alumina particles cannot be predicted with the model. Rather, the average sulphate and Al_2O_3 composition fractions of all the aerosols in a given model size bin are predicted.

The maximum concentration of (nearly) pure Al_2O_3 particles identifiable among the total population of aerosols in size bin i was first estimated as

$$n_i(\text{Al})_m = n_i f_i(\text{Al}) \quad (1)$$

where n_i is the concentration of aerosols of size i (number cm^{-3} , excluding CN), and $f_i(\text{Al})$ is the volume fraction of Al_2O_3 in aerosols of size i . However, the present model also provides information on the aerosol core size distribution in terms of the core volume moment fraction, h_i , which is a measure of the size dispersion of the cores in aerosols of size i , independent of core composition¹⁴. It can be shown that using h_i and f_i [where $f_i = \sum_j f_i(j)$], the maximum number of essentially-pure core aerosols may also be approximated as

$$n_i(\text{core})_m \doteq n_i [h_i^2 - f_i^2] / [1 + h_i^2 - 2f_i] \quad (2)$$

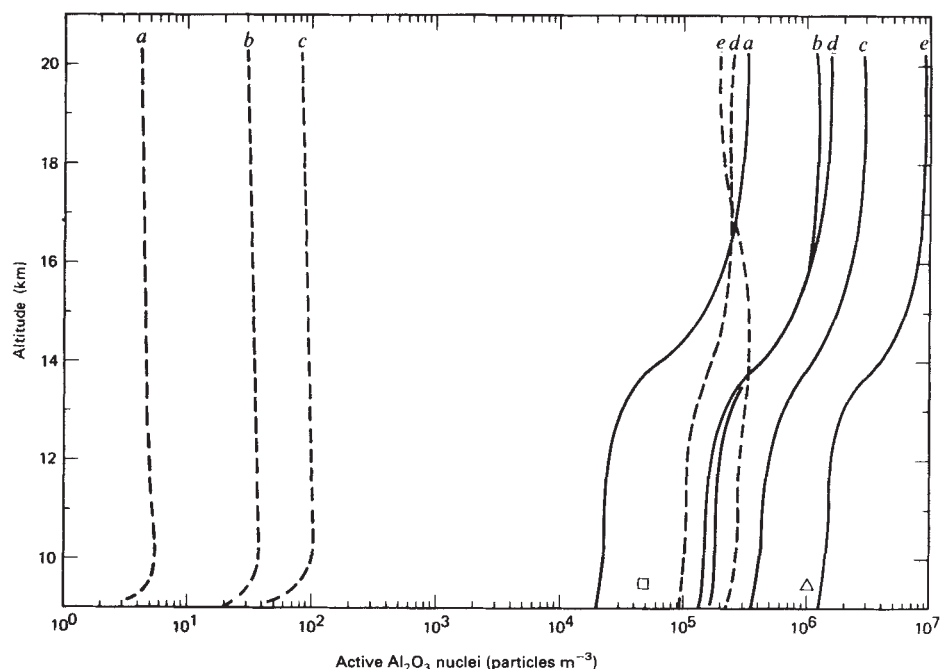
Accordingly, the maximum concentration of aerosol SSIN of size i was taken as the lesser of $n_i(\text{Al})_m$ and $n_i(\text{core})_m$.

The 'average' number of aerosol SSIN was calculated by assuming that all of the aerosols of size i have an average coating, or film, of contaminating material (mainly liquid sulphuric acid). If, in any size bin, the contamination amounted to more than a uniform monolayer coating $\approx 4\text{ \AA}$ thick, all of the aerosols in that bin were considered to be poisoned.

The effect of particle size on the activity of rocket nuclei was estimated in two ways. In one case, the SSIN concentrations deduced above (including the contributions of CN and droplets) were truncated to exclude particles $<0.01\text{ }\mu\text{m}$ radius; such particles generally require supersaturations exceeding $\sim 10\%$ to cause nucleation at $\sim -20^\circ$ to -30°C (refs 16, 17). In the second approach, the effect of particle size on the water vapour pressures of small ice spheres was calculated using the Kelvin relation¹⁷. At a temperature of -30°C (adopted to estimate the critical size at each altitude) the minimum nuclei dimensions corresponding to water vapor supersaturations of 1, 10, 50, 100 and 900% are 0.19, 0.020, 0.0046, 0.0027 and 0.0008- μm radius, respectively.

Figure 1 shows projected aerosol size distributions at 20 km for several different Al_2O_3 emission size spectra, $r^{-\alpha}$. In the size range 1–10 μm , the aerosols are essentially pure alumina ($>95\%$ pure). The data of Brownlee *et al.*³ are plotted for comparison. The close agreement between the observed size distribution and the computational results suggests that the present shuttle simulations are physically reasonable. Brownlee *et al.*³ estimated the global stratospheric burden of Al_2O_3 in 1974–5 as ~ 100 tonnes. The stratospheric lifetime of particles in the size range 1–5 μm is ~ 1 month, implying an Al_2O_3 mass input rate of $\sim 1,000$ tonnes yr^{-1} worldwide. The space shuttle calculations in Fig. 1 yield ~ 10 times as many rocket alumina particles for a hemispherical input rate of $\sim 5,000$ tonnes yr^{-1} , in clear accord with the Brownlee *et al.* observations. The

Fig. 3 Predicted space shuttle ice nuclei (SSIN) concentrations in the lower stratosphere and upper troposphere. The cases shown are: a, $r^{-3.5}$, $\tau_n = 100\text{ s}$, $\tau_r = 10^6\text{ s}$; b, r^{-4} , $\tau_n = 100\text{ s}$, $\tau_r = 10^6\text{ s}$; c, $r^{-4.5}$, $\tau_n = 100\text{ s}$, $\tau_r = 10^6\text{ s}$; d, r^{-4} , $\tau_n = 10^6\text{ s}$, $\tau_r = 10^7\text{ s}$; e, r^{-4} , $\tau_n = 10^6\text{ s}$, $\tau_r = 10^6\text{ s}$ (see the text for an explanation of these parameters). In cases a–d, 52 shuttle launches per year were assumed, with the effluent spread uniformly over the Northern Hemisphere. In case e, a 10-fold increase in the numbers of launches was assumed. A minimum nuclei radius of 0.01 μm was used to estimate the SSIN abundances. Maximum (solid line) and average (dashed line) numbers of Al_2O_3 nuclei are shown. Δ , Observed ice nuclei concentration in cirrus clouds at -30°C . $\square$, Maximum ice crystal concentration in cirrus clouds.



corresponding predicted global mass of the stratospheric alumina layer is $\sim 1,000$ tonnes.

The computed size distributions in Fig. 1 have a relatively simple explanation. If rocket particles are injected at all altitudes with the same size distribution ($r^{-\alpha}$), and if particle growth by condensation and coagulation can be neglected, then the resultant atmospheric aerosol size distribution depends on radius roughly as $r^{-\alpha-\beta}$, where β is the exponent of the radius dependence of the particle fall velocity; that is, $v(r) \propto r^\beta$. In the lower stratosphere, $1 < \beta < 2$ for particles with radii in the size range 1–10 μm . Thus, for Al_2O_3 rocket particles one expects $5 \leq (\alpha + \beta) \leq 6$, as is found in Fig. 1.

Figure 2 illustrates the influence of water vapour supersaturation on apparent SSIN concentrations. As expected, higher supersaturations yield larger IN abundances. The abrupt increase in the 'average' IN curves at S values of 40–80% is due to a substantial number of nearly-pure rocket dust particles at sizes below $\sim 0.01 \mu\text{m}$. For meteorologically significant supersaturations ($< 10\%$), the average number of IN in Fig. 2 is nearly four orders of magnitude smaller than the maximum number of IN. The difference represents, in one sense, the uncertainty in predictions of rocket IN abundances.

Predicted SSIN concentrations in the upper troposphere and lower stratosphere, for various physical parameter values, are given in Fig. 3. Several trends in the data are noteworthy. As in Fig. 2, with fast Al_2O_3 nucleation ($\tau_n = 100$ s), there is a difference of a factor of 10^4 between the 'average' and 'maximum' numbers of SSIN. Obviously, there is ample sulphate material present in the atmosphere to contaminate most of the injected dust grains. The maximum SSIN concentrations would apply if Al_2O_3 particle surfaces were not efficiently nucleated and poisoned by H_2SO_4 vapour (see, for example, cases *d* and *e* in Fig. 3). Comparison of cases *b*, *d* and *e* suggests that the average SSIN concentration is roughly proportional to τ_n^{-1} . In other words, once nucleated, most Al_2O_3 grains accumulate enough sulphur vapour to be poisoned (in the model). Rainout has a much smaller influence on SSIN concentrations; the residence time of injected particles in the upper troposphere is controlled primarily by vertical diffusion (for the cases studied).

One might expect a short nucleation time ($\tau_n \sim 100$ s) to apply to Al_2O_3 dust released in the upper atmosphere. For one thing, the H_2SO_4 supersaturation ratio between 10 and 20 km altitude is quite large, $\sim 10^4$ – 10^5 (ref. 16). Hence, adsorption of H_2SO_4 on alumina surfaces should commence immediately on injection. Continuing growth of the dust particles larger than the critical (Kelvin) size would also be guaranteed. However, a factor which causes the effective nucleation time to be much longer is the time required to deposit a monolayer of acid molecules on the dust grain surfaces. The ambient quantity of H_2SO_4 vapour available in the neighbourhood of the launch plume is insufficient to coat more than a few per cent of the rocket particles. Nevertheless, it may be demonstrated that, within a day or so, individual launch plumes are dispersed widely enough for extensive acid adsorption to occur. This does not take into consideration the HCl vapour in the early launch plume, which acts to poison the dust surfaces at high concentrations^{9–11}, but appears to activate the surfaces at low concentrations¹³. Accordingly, an effective dust nucleation time $\leq 10^5$ s appears to be most appropriate for stratospheric conditions.

Many of the SSIN residing in the upper troposphere are injected between 8 and 12 km. At these altitudes, the transient space shuttle launch clouds should be considered. Case *e* in Fig. 3, which corresponds to a 10-fold enhancement in the average shuttle particle deposition rate, might crudely forecast the elevated IN concentrations to be expected in isolated launch plumes.

The maximum number of rocket IN that could be generated from 1 g of fuel, based on an r^{-4} Al_2O_3 dust size distribution and a minimum activation radius of 0.01 μm , is $\sim 10^{15}$ IN g^{-1} . This number corresponds closely to the maximum number predicted by the model. By contrast, Hindman *et al.*¹² measured

$\sim 10^{10}$ IN g^{-1} in small shuttle propellant burns, while the present calculations yield estimates of average long-term shuttle IN yields $\sim 10^{11}$ IN g^{-1} . Hence, the maximum SSIN profiles in Fig. 3 are very likely to represent upper-limit global-scale concentrations of artificial ice nuclei produced by space shuttles.

Typical IN concentrations measured at cirrus cloud heights are indicated in Fig. 3 (ref. 17). Maximum concentrations of ice crystals observed in cirrus clouds are also shown. Based on a comparison of measured and predicted IN concentrations, and the previous discussion, we conclude that planned space shuttle activity probably will not increase the global abundance of ice nuclei at cirrus cloud levels by more than a factor of 2, although larger transient increases might occur locally. Injected space shuttle alumina particles, on the other hand, could enhance the concentrations of stratospheric aerosols in the size range 1–10 μm radius by a factor of 10 or more, creating in the stratosphere an aluminium oxide layer exceeding 1,000 tonnes in weight.

Received 1 April; accepted 30 June 1982.

1. Potter, A. E. *Proc. Space Shuttle Environmental Assessment Workshop on Stratospheric Effects* (NASA TM X-58198, Houston, Texas, 1977).
2. Pollack, J. B. *et al. J. geophys. Res.* **81**, 1071–1083 (1976).
3. Brownlee, D. E., Ferry, G. V. & Tomandl, D. *Science* **191**, 1270–1271 (1976).
4. Hofmann, D. J., Carroll, D. E. & Rosen, J. M. *Geophys. Res. Lett.* **2**, 113–116 (1975).
5. Pollack, J. B., Toon, O. B., Summers, A., Van Camp, W. & Baldwin, B. *J. appl. Met.* **15**, 247–258 (1976).
6. Turco, R. P. *et al. J. appl. Met.* **19**, 78–89 (1980).
7. Hobbs, P. V. *Ice Physics* (Oxford University Press, 1974).
8. Parungo, F. P. & Allee, P. A. *J. appl. Met.* **17**, 1856–1863 (1978).
9. Parungo, F. P. & Allee, P. A. *J. appl. Met.* **19**, 128–130 (1980).
10. Hindman, E. E., II, & Lala, G. G. *J. appl. Met.* **19**, 122–128 (1980).
11. Hindman, E. E., Radke, L. F. & Eltgroth, M. W. *J. appl. Met.* (submitted).
12. Hindman, E. E., II, Garvey, D. M., Langer, G., Odencrantz, F. K. & Gregory, G. L. *J. appl. Met.* **19**, 175–184 (1980).
13. Hindman, E. E. & Finnegan, W. G. *Pap. 8th Conf. on Inadvertent and Planned Weather Modification*, Reno (1981).
14. Turco, R. P., Hamill, P., Toon, O. B., Whitten, R. C. & Kiang, C. S. *J. Atmos. Sci.* **36**, 699–717 (1979); *The NASA-Ames Research Center Stratospheric Aerosol Model I. Physical Processes and Computational Analogs* (NASA TP 1362, 1979).
15. Turco, R. P., Toon, O. B., Hamill, P. & Whitten, R. C. *J. geophys. Res.* **86**, 1113–1128 (1981).
16. Hamill, P., Turco, R. P., Toon, O. B., Kiang, C. S. & Whitten, R. C. *J. Aerosol Sci.* (in the press).
17. Pruppacher, H. R. & Klett, J. D. *Microphysics of Clouds and Precipitation* Ch. 9 (Reidel, Dordrecht, 1978).

The Earth's early hydrosphere

A. Henderson-Sellers

Department of Geography, University of Liverpool, PO Box 147, Liverpool L69 3BX, UK

J. Graham Cogley

Department of Geography, Trent University, Peterborough, Ontario, Canada K9J 7B8

The date at which the Earth's hydrosphere came into existence is unknown. Although all current geological models for the very early Precambrian predict or suppose a hydrosphere^{1–4}, there are difficulties in reconciling this with geological and astronomical theories of planetary formation. Furthermore the effect of a substantial hydrosphere on the Precambrian climatic environment is difficult to determine. Recent developments make it possible to synthesize a picture of this environment during the 700 Myr preceding the deposition of the oldest known sediments⁵. We attempt such a synthesis here, drawing together work on the dynamics of accretion from the proto-planetary nebula^{6,7} and of planetary differentiation⁷, on the role of volatiles in impact cratering^{8–10}, and on the composition^{11–22}, isotope systematics^{23–25} and radiation climatology^{26–29} of the early hydrosphere. We can reconcile these diverse contributions if the hydrosphere—the inventory of excess volatiles¹—is allowed to differentiate during accretion, with most of the H_2O and CO_2 going rapidly into oceans and sediments. Both gradual outgassing³⁰ and a massive CO_2 atmosphere² are unlikely, and a reduced mantle beneath a neutral hydrosphere is not paradoxical^{1,3,17}. Results from both a one-dimensional radiative-convective (1D RC) model and a general circulation model suggest that self-regulating mechanisms are important in the climate of this early hydrosphere.

The idea of early degassing³¹ now seems more probable than slower degassing. The possible histories which cannot yet be excluded are very diverse, but the Earth probably grew over 10^7 yr or more⁶ from a population of planetesimals, the largest having a mass not less than 1/500 that of the total accreted mass⁷. Not all of the impact heat is radiated away. Some of it is buried, and it sets up convection in the outer accreting layer and triggers formation of the core, perhaps as early as accreted mass $\sim 0.1 \times$ final mass. The energy released by core formation⁷, and by subsequent steady settling of dense phases, establishes convection throughout the growing mantle, and volatiles differentiate upwards as iron phases segregate downwards. A growing planet tends to 'reject' icy impactors^{8,9}. That is, the ratio r (ejecta mass escaping the planet/impactor mass) is large when the impactor density is small, and metals and silicates accrete much more readily than do ices; r depends also on impact velocity, however, and during growth the velocity needed for net erosion ($r > 1$) increases. In the late stages of growth only the most energetic impactors actually sputter the surface, and because the planet has depleted its source of planetesimals the latest arrivals are previously-rejected, volatile-rich bodies. This inhomogeneous accretion³² yields an outer layer of more or less volatile-charged silicates, which separates efficiently into veneer and substrate under the influence of repeated impacts. Vapour volumes exceeding impactor volume are produced in a large proportion of impacts^{8,10}. Hydrous and carbonate vapours disperse more widely, and precipitate less completely, than silicate vapours¹⁰.

The earliest mantle, with volatiles strongly concentrated upwards, was thus overlain by an impact-processed protocrust, hydrated and carbonated but of uncertain thickness. Two points deserve emphasis:

First, geothermal energy sources were much larger during accretion than at any later time; models with outgassing due to radioactive warmup³⁰, or similar gradual processes, now seem mechanically and thermally improbable, and it is safest to conclude that the excess volatiles had a mass comparable with that of the present by the time accretion was completed.

Second, enrichment in ^{129}Xe of Earth-interior gases with respect to the present-day atmosphere and to Archaean shales^{23,24} requires outgassing of the atmosphere within a few half-lives (17 Myr) of ^{129}I , with negligible net additions since 3,300 Myr ago²⁵. The history described above explains this important observation well.

It is important to consider the oxidation state of the early hydrosphere. It seems to be accepted that H_2O and CO_2 were significant components even of the earliest atmosphere¹¹, but questions persist over the reduced species H_2 , NH_3 and CH_4 (refs 2, 3, 12).

Modern volcanic gases have elemental bulk ratios similar to those of the excess volatiles¹³, and have high-temperature low-pressure equilibria^{14,15}; they seem to sample the modern hydrosphere, but not necessarily the juvenile source during accretion. The lower continental crust seems to be anhydrous¹⁶, its volatiles dominated by CO_2 . This is consistent with empirical thermodynamics¹⁷, as is the observation that the upper mantle is strongly reduced^{3,18,19}. The modern mantle seems to be closer to equilibrium with the core than with the hydrosphere. This pattern would evolve from our picture of accretion if (1) some of the H_2 in a reduced proto-mantle went to the core²⁰ and/or (2) the primitive outgassed volatiles were reduced. In case (2) we must explain how and when the hydrosphere lost its evolved reduced gases.

The primary control of atmospheric H_2 is escape from the top of the atmosphere²¹, which for exospheric temperatures $T_x > 600$ K is controlled by hydrogen diffusion into the upper atmosphere. Most OH and H radicals, from photolysis of H_2O for example, will recombine, and only a fraction of the hydrogen arriving from below actually escapes (as H). Likely values²² of T_x exceed 1,300 K, a temperature at which, if diffusion into the upper atmosphere were not restricted, the early Earth could evaporate an excess of hydrogen rather efficiently. In an early

atmosphere without ozone, a very cold and nearly isothermal stratosphere³³ might indeed allow more H_2O vapour to diffuse upwards as the temperature inversion at the tropopause would be absent. A great excess of hydrogen, if produced by a large-body impact, would blow away very rapidly, possibly in a few minutes²¹.

Other reduced constituents are also unlikely to persist. Ammonia is highly soluble in water and would be removed rapidly by rainout and by solution at the ocean surface. It would also be destroyed by photolysis²⁶. It has been calculated that a mixing ratio of $< 10^{-4}$ of NH_3 will convert irreversibly to N_2 in < 40 yr (ref. 27). Methane photolyses at wavelengths < 160 nm, where the incident solar flux is two orders of magnitude less than in the region important for the destruction of NH_3 ($\lambda < 230$ nm), and where atmospheric water vapour will provide a substantial shield²⁸. CH_4 is rapidly destroyed above 100 km (ref. 29), but at the surface it appears much more stable. However, photolysis of H_2O vapour forms large quantities of OH, an important intermediary in the methane oxidation chain. At an OH concentration of 10^5 cm^{-3} in the early troposphere²⁸, the life time of CH_4 at the surface is < 50 yr.

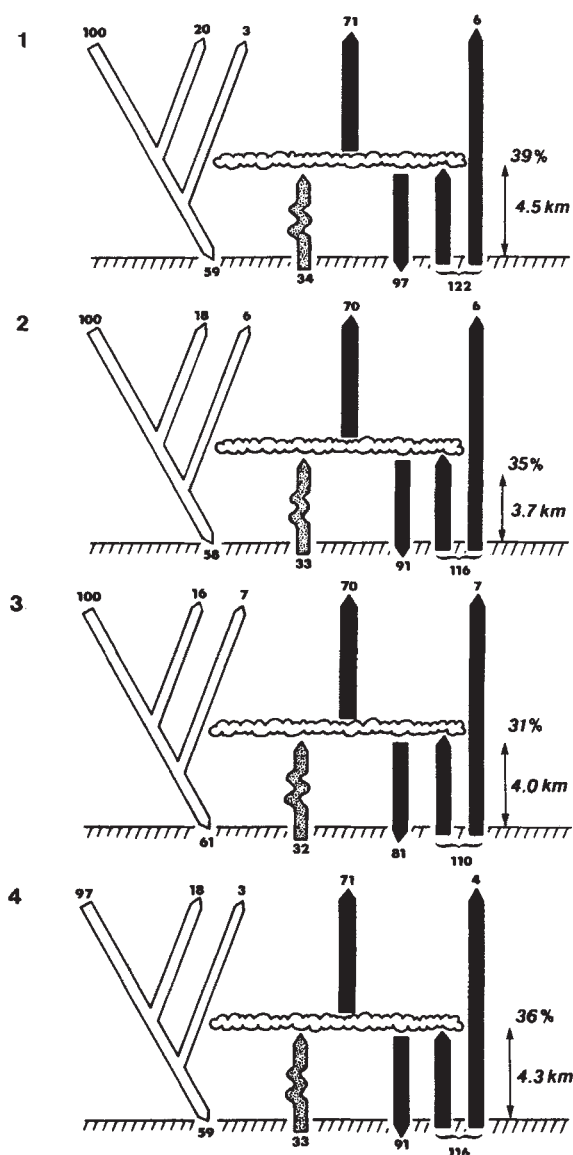


Fig. 1 Four likely configurations of the atmosphere/hydrosphere of the early Earth. The calculated fluxes have been normalized to the incident solar radiation of the standard case (1) (100 units = $\frac{1}{4}$ of $1,100 \text{ W m}^{-2}$). Solar fluxes (open arrows) are on the left of each sketch; upward arrows represent atmosphere (left) and surface (right) reflection. The wavy arrow represents convective heat flux from the surface, and infrared fluxes are on the right (solid arrows). Note the variations in mean cloud cover (per cent) and cloud height (km) at right.

Table 1 Radiative-convective model results

	Case 1	Case 2	Case 3	Case 4
Prescribed				
Incident solar radiation (W m^{-2})	1,100	1,100	1,100	1,067
Surface albedo	0.05	0.10	0.10	0.05
CO ₂ (p.p.m.v.)	1,650	1,650	330	1,650
Results				
T_s (K)	277.0	273.8	270.2	274.1
Cloud cover (%)	39.0	34.8	31.3	35.8
Mean cloud height (km)	4.48	3.71	3.99	4.32
Planetary albedo	0.23	0.24	0.23	0.22
Greenhouse ΔT (K)	29.7	27.7	22.9	28.3

The greenhouse temperature increment (present-day value ~ 33 K) is $T_s - [S(1-A)/4\sigma]^{1/4}$, where σ is the Stefan-Boltzmann constant, A is the planetary albedo and T_s is the mean surface temperature.

Hence in a water-vapour-dominated atmosphere neither NH_3 nor CH_4 can persist, H_2 escapes rapidly, and a reduced-gas greenhouse is at most transient, dissipating on time-scales shorter than that of accretion. Greenhouse warming was controlled by the amounts of CO_2 and H_2O vapour in the early atmosphere.

We assume that there is always enough greenhouse gas to avoid the very stable³⁴ 'deep freeze' predicted by some models of the present-day climate. Note, however, that little reliance can be placed on predictions made by EBMs³⁴ unless careful consideration is made of the changes in the horizontal advection term and the effect of clouds on the planetary albedo. We assume further that surface albedo was controlled by the presence of liquid water, as exposed land was probably less extensive (0–10% of total planetary surface) than during the Phanerozoic (20–30%)³⁵. The earliest topography was probably unimodal, with a few kilometres of relief on a mechanically weak proto-crust shaped by large impacts. The continental mass has increased in the course of time^{36,37}, and high-standing continents probably did not appear until after the end of accretion.

A massive CO_2 atmosphere, $P_{\text{CO}_2} \approx 1,000$ mbar, is unlikely. It is not necessary in our model to invoke a passage of CO_2 from atmosphere to sediments on the long time scale of continental differentiation⁵, because carbonates would form in the aftermath of large-body impacts. In any case, the P_{CO_2} of a surface mixed layer (say 50m deep and covering 70–100% of the planet) would equilibrate with the atmosphere in a few months or years. CO_2 downwelling, and reaction with the sea floor at geohydrological (a few times 10^4 yr)³⁸ and geological ($\approx 10^6$ – 10^7 yr) rates, would increase with atmospheric CO_2 . Massive quantities of both liquid water and gaseous CO_2 are thus not compatible for more than short times. We choose $\text{CO}_2 = 1,650$ p.p.m.v. ($5 \times \text{PAL}$) in the work described below. Actual concentrations might have been much higher. However, our results (Table 1) suggest that if they were of the order of 300,000 p.p.m.v. (refs 2, 11) the Earth could have evolved into a runaway greenhouse. The hydrospheric feedbacks we describe give temperatures warmer than those of a non-interactive model both for reduced solar flux and for increased levels of greenhouse gases³³. Some early, non-hydrothermal³⁹ sediments do suggest deposition from warm water⁴⁰, but there is no evidence for a runaway greenhouse in the geological or geochemical records.

Atmospheric water vapour concentration depends primarily on T_s . Cloud effects have generally been simulated crudely or neglected in studies of the early atmosphere^{2,11,41} and this may⁴² be a serious omission. We explore cloud effects with the GISS 1D RC model⁴³ due to Lacis, incorporating an interactive cloud prediction scheme³³ which makes the model much less sensitive: $\beta < \approx 80$, down from $\beta \approx 100$ as found in prescribed-cloud models⁴⁴. ($\beta = S\Delta T/\Delta S$, where S is the present-day solar constant.) At each altitude, cloud cover depends on the water mixing ratio which is calculated from the mixing ratio at the

previous time step by accounting for additions from condensation and removal by rainfall. In common with most 1D RC models a fixed lapse rate of 6.5 K km^{-1} and fixed relative humidity are assumed. Additionally the Bowen ratio linking latent heat flux to convective flux is a function of T_s but is assumed to be constant with height. When clouds are predicted at levels with $T < 246$ K they are assumed to be ice rather than water clouds and the effective particle radius and optical depth are adjusted accordingly. Since the cloud cover is predicted by the model, the feedbacks are internally consistent. In particular, an increase in high, optically-thin clouds tends to feed back positively on temperature³³.

Our standard state is shown as case (1) in Fig. 1 and Table 1. The solar flux is $1,100 \text{ W m}^{-2}$ (80% of the present-day flux⁴⁵), and the albedo and atmospheric CO_2 are consistent with our model for the close of accretion. T_s is 277 K, warm enough to support liquid water. The dark surface, and IR absorbers in the atmosphere, contribute to a greenhouse effect almost as large as it is at present. The albedo is rather low; if we increase it to represent a substantial emergence of land (case (2)), T_s drops to 274 K. If we next reduce CO_2 to its modern value (case (3)), perhaps representing consumption of CO_2 in silicate weathering, the greenhouse is weaker (23 K), although even now at $T_s = 270$ K the oceans will probably not freeze over. A largely ocean-covered Earth^{36,37} is extremely difficult to glaci-ate. The pole-equator temperature gradient is likely to be smaller than today; but even with temperatures significantly < 273 K at the poles, the absence of large land masses either to be the site of glaciers or to constrain oceanic circulation is likely to inhibit severely ice formation.

Note that the calculated temperatures (cases (1)–(3)) probably represent the lower limit to the evolutionary curve of T_s . This is because we have chosen to include the smallest likely CO_2 concentrations to focus attention on the feedback effects operating as a result of the existence of the hydrosphere. Despite this minimization of the CO_2 greenhouse effect, the lower limit temperature curve presented here is physically reasonable and is consistent with the geological data. This is because the sensitivity, β , of the interactive-cloud 1D RC model³³ is lower than that of any other climate model. If cases (1), (2) and (3) followed each other, as they well might have, then glaciation becomes more likely with time. More importantly, however, the interactive-cloud model continues to predict a stable climatic régime in the face of these radical climatic perturbations.

The most dramatic single perturbation of early climate is probably the large impact. Mass input rates were negligible in comparison, but such impacts certainly continued after the end of accretion; the lunar bombardment which formed Imbrium at 4,000–3,800 Myr may have obliterated evidence of still earlier bombardment⁴⁶. The most significant atmospheric effects would probably be changes in the vertical distribution of water vapour and liquid water. The climatic perturbation due to a 10^{15} kg impactor depositing $\approx 10^{16}$ kg of water in the upper atmosphere (only 0.2×10^{16} kg below 10 km) has recently been calculated using the GISS general circulation climate model^{47,48}. Changes were transient (with a decay time < 1 yr) and small (temperature increasing by a maximum of 0.5 K after 5 months). Long-term effects are seen only if perturbations to the lower atmosphere are strengthened by continual replenishment of species^{49,50}; positive feedbacks internal to the climatic system do not seem likely.

Placement of the vapour and debris at a much higher level in the atmosphere, it has been suggested⁵¹, may create a permanent near-worldwide layer of mesospheric ice clouds, cooling the surface through a reduction of planetary albedo by several per cent. We can simulate this albedo effect by reducing solar flux to $1,067 \text{ W m}^{-2}$ (case (4) of Table 1 and Fig. 1). The result is a drop of only 3 K in T_s , to 274 K, within the range of cases (1)–(3) and unlikely to lead to a global freeze-over.

Comparable remarks hold for other perturbations. Even during accretion, interior convection can dissipate the input of

impact and gravitational heat if lava lakes cover a fraction (10^{-3} – 10^{-4}) of the surface⁷. Shallow heat radiates away rapidly from vapour clouds, ejecta and melt pools. Buried heat enhances⁵² a geothermal flux due to radioactivity and energy of accretion, but even with extreme assumptions⁵³ this flux does not reach 1 W m^{-2} . Short-term changes with complex regional variability will follow large impacts, but we have no evidence that the climate cannot recover from these sporadic intervals of disturbance.

A substantial early hydrosphere, then, is both consistent with results from a variety of disciplines and, as our results show, powerful in solving the paradox of the faint young Sun⁴⁵. The most plausible date for its origin is at the very beginning. Most of its properties, and especially its stability against large, irreversible changes of composition and climate, seem to follow if we identify the principal surface volatile as liquid water. There was thus a long expanse of pre-Isuan time during which the Earth could have sheltered and nurtured organisms or their precursors, and this expanse of time would surely repay closer attention.

We thank W. Rossow for helpful discussions in this work, which was begun while A. H.-S. was a NRC (US) visiting research fellow at the Goddard Institute for Space Studies. J. G. C. thanks NSERC (Canada) for support.

Received 8 February; accepted 23 June 1982.

- Walker, J. C. G. *Evolution of the Atmosphere* (Macmillan, New York, 1977).
- Owen, T., Cess, R. D. & Ramanathan, V. *Nature* **277**, 640–642 (1979).
- Arculus, R. J. & Delano, J. W. *Nature* **288**, 72–74 (1980).
- Henderson-Sellers, A., Benlow, A. & Meadows, A. J. *Q. J. R. astr. Soc.* **21**, 74–81 (1980).
- Moorbath, S. *Geol. Ass. Can. Spec. pap.* **20**, 89–115 (1981).
- Safronov, V. S. *Icarus* **33**, 3–12 (1978); *Evolution of the Protoplanetary Cloud and Formation of the Earth and the Planets* (NASA Tech. Transl. TTF-677 Washington DC, 1972).
- Kaula, W. M. *Geol. Ass. Can. Spec. pap.* **20**, 25–34 (1981); *J. geophys. Res.* **84**, 999–1008 (1979).
- O'Keefe, J. D. & Ahrens, T. J. *J. geophys. Res.* (submitted); *Proc. 8th Lunar Sci. Conf.* 3357–3374 (1977).
- Orpahl, D. L., Roddy, D. J., Schultz, P. H., Borden, W. F. & Larson, S. A. *Pap. Conf. on Large Body Impacts and Terrestrial Evolution*, Snowbird, Utah, 41 (Lunar Planetary Institute, Houston, 1981).
- Kieffer, S. W. & Simonds, C. H. *Rev. geophys. Space Phys.* **18**, 143–181 (1980).
- Hart, M. H. *Icarus* **33**, 23–39 (1978).
- Henderson-Sellers, A. & Schwartz, A. W. *Nature* **287**, 526–528 (1980).
- Anderson, A. T. *Rev. geophys. Space Phys.* **13**, 37–55 (1975).
- Fudali, R. F. *Geochim. cosmochim. Acta* **29**, 1063–1075 (1965).
- Heald, E. F., Naughton, J. J. & Barnes, I. L. *J. geophys. Res.* **68**, 545–557 (1963).
- Newton, R. C., Smith, J. V. & Windley, B. F. *Nature* **288**, 45–50 (1980).
- Perchuk, L. L. in *Energetics of Geological Processes* (eds Saxena, S. K. & Bhattacharji, S.) 285–352 (Springer, New York, 1977).
- Melton, C. E. & Giardini, A. A. *Am. Miner.* **66**, 746–750 (1981).
- Kramers, J. D., Smith, C. B., Lock, N. P., Harmon, R. S. & F. R. Boyd *Nature* **291**, 53–56 (1981).
- Stevenson, D. J. *Nature* **268**, 130–131 (1977).
- Hunten, D. M. *J. atmos. Sci.* **30**, 1481–1494 (1973).
- Shimizu, M. *Precamb. Res.* **3**, 463–470 (1976).
- Frick, U. & Chang, S. *Proc. 8th Lunar Sci. Conf.* 263–272 (1977).
- Phinney, D. *Earth planet. Sci. Lett.* **16**, 413–420 (1972).
- Thomsen, L. J. *geophys. Res.* **85**, 4374–4378 (1980).
- Ferris, J. P. & Nicodem, D. E. *Nature* **238**, 268–269 (1972).
- Kuhn, W. R. & Atreya, S. K. *Icarus* **37**, 207–213 (1979).
- Levine, J. S. & Augustsson, T. R. in *Proc. College Park Colloquia on the Voyager Mission: Implications for Planetary Biology*, Oct. 4–6 (1981).
- Lasaga, A. C., Holland, H. D. & Dwyer, H. D. *Science* **174**, 53–55 (1971).
- Li, Y.-H. *Am. J. Sci.* **272**, 119–137 (1972).
- Fanale, F. P. *Chem. Geol.* **8**, 79–105 (1977).
- Anders, E. & Owen, T. *Science* **198**, 453–465 (1977).
- Rossow, W. B., Henderson-Sellers, A. & Weinreich, S. K. *Science* (in the press).
- Ghil, M. *J. atmos. Sci.* **33**, 3–20 (1976).
- Cogley, J. G. *Nature* **291**, 56–58 (1981).
- McCulloch, M. T. & Wasserburg, G. J. *Science* **200**, 1003–1011 (1978).
- Veizer, J. & Jansen, S. L. *J. Geol.* **87**, 341–370 (1979).
- Gregory, R. T. & Taylor, H. P. *J. geophys. Res.* **86**, 2737–2755 (1981).
- Knauth, L. P. & Epstein, S. *Geochim. cosmochim. Acta* **40**, 1095–1108 (1976).
- Costa, U. R., Fyfe, W. S., Nesbitt, H. W. & Kerrich, R. *EOS* **61**, 386 (1980).
- Sagan, C. & Mullen, G. *Science* **177**, 52–56 (1972).
- Wang, W.-C., Rossow, W. B., Yao, M.-S. & Wolfson, M. J. *J. atmos. Sci.* **38**, 1167–1178 (1981).
- Hansen, J. E. *et al. Science* **213**, 957–966 (1981).
- Wang, W.-C. & Stone, P. H. *J. atmos. Sci.* **37**, 545–552 (1980).
- Newman, M. J. & Rood, R. T. *Science* **198**, 1035–1037 (1977).
- Wasserburg, G. J., Papanastassiou, D. A., Tera, F. & Huneke, J. C. *Phil. Trans. R. Soc. A* **285**, 7–22 (1977).
- Kuhn, J. in *Proc. GISS Summer Inst. Planet. Climate*, 1981 (NASA, New York, 1981).
- Hansen, J. E., Russell, G., Rind, D., Stone, P. H., Lacis, A., Lebedeff, S., Ruedy, R. & Travis, L. *Mon. Weath. Rev.* (submitted).
- O'Keefe, J. D. & Ahrens, T. J. *Pap. Conf. on Large Body Impacts and Terrestrial Evolution*, Snowbird, Utah, 39 (Lunar Planetary Institute, Houston, 1981).
- Toon, O. B. *et al. EOS* **62**, 886 (1981).
- McKay, C. P. & Thomas, G. E. *Pap. Conf. on Large Body Impacts and Terrestrial Evolution*, Snowbird, Utah, 33 (Lunar Planetary Institute, Houston, 1981).
- Onorato, P. I. K., Uhlmann, D. R. & Simonds, C. H. *J. geophys. Res.* **83**, 2789–2798 (1978).
- Davies, G. F. *J. geophys. Res.* **85**, 2517–2530 (1980).

Optimized anomalous dispersion in crystallography: a synchrotron X-ray polychromatic simultaneous profile method

U. W. Arndt*, T. J. Greenhough†, J. R. Helliwell†‡, J. A. K. Howard§, S. A. Rule† & A. W. Thompson‡

* MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

† Department of Physics, University of Keele, Keele, Staffordshire ST5 5BG, UK

‡ SERC, Daresbury Laboratory, Daresbury, Warrington WA4 4AD, UK

§ School of Chemistry, University of Bristol, Cantock's Close, Bristol BS8 1TS, UK

In principle, it is possible to phase the X-ray reflections from a macromolecular crystal which contains one heavy atom per molecule, without having recourse to the conventional method of multiple isomorphous replacement, by making use of anomalous dispersion effects alone¹. To do this, it is necessary to measure each reflection over a narrow range of wavelengths centred on an absorption edge of the heavy atom, where f' and f'' , the real and imaginary anomalous components of the scattering factor, are varying rapidly. The availability of powerful synchrotron radiation sources with a continuous energy spectrum has stimulated interest in such phasing methods, but previously experiments¹ have been limited to making measurements on only one reflection at a time and/or to measuring at only a few discrete wavelengths near the absorption edge. We are developing a method at the Daresbury Synchrotron Radiation Source (SRS) which produces an energy profile along an axis of each and every diffraction spot in a screenless oscillation photograph. We show here that the approach is feasible using a single crystal of heptahydrido bis(diisopropylphenyl)phosphine rhenium using the L_{III} absorption edge of rhenium at a wavelength of $1.1772 \text{ \AA}$.

The white synchrotron radiation beam from a storage ring diverges from a small region at the tangent point of the beam line. A narrow wavelength band can be selected from this beam by reflection from a crystal monochromator; a cylindrically curved monochromator produces a converging X-ray beam. At the SRS protein crystallography station² the monochromator is a perfect crystal of either Si or Ge with the (111) planes cut at an angle α to the surface. The triangular crystal is clamped at its base and deflected at its tip: in this way^{2–5} an accurately cylindrical curvature is achieved.

There is one curvature of the monochromator for a given Bragg angle (θ), angle of cut (α), image (p') and object (p) distances where all rays which are brought to a focus have been reflected at an equal angle so that they are all of the same wavelength. This 'achromatic condition' is usually referred to as the 'Guinier position' of the monochromator. However, for monochromator curvatures, R , other than that for the 'Guinier position', rays from a point source make different angles of incidence along the length, L , of the monochromator, over a range $\delta\theta_{\text{corr}}$ where^{2,4–6}

$$\delta\theta_{\text{corr}} = \frac{L}{2} \left\{ \frac{\sin(\theta - \alpha)}{p'} - \frac{\sin(\theta + \alpha)}{p} \right\}$$

$$\left| \left(\frac{\delta E}{E} \right) \right|_{\text{corr}} = \left(\frac{\delta \lambda}{\lambda} \right)_{\text{corr}} \approx \delta\theta_{\text{corr}} \cot \theta$$

and

$$\frac{2}{R} = \frac{\sin(\theta - \alpha)}{p'} + \frac{\sin(\theta + \alpha)}{p}$$

(At the Guinier position $(\delta\lambda/\lambda)_{\text{corr}} = 0 = \delta\theta_{\text{corr}}$ and

$$\frac{1}{R} = \sin \frac{(\theta - \alpha)}{p'} = \sin \frac{(\theta + \alpha)}{p}.)$$

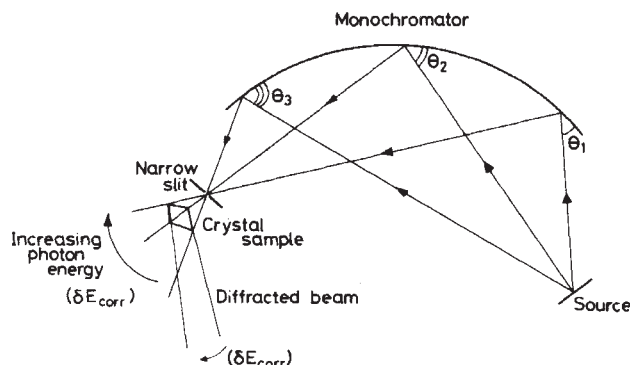


Fig. 1 When the monochromator curvature is not at the 'Guinier position' rays from a point on the source are incident at different angles to different points along the length of the monochromator and are reflected at different wavelengths (energies). At a point in the focus, defined by the narrow slit, the converging rays possess an exact correlation of photon energy with direction. The particular curvature of the monochromator that is needed is that for which $(\theta_3 - \theta_1) = \delta\theta_{\text{corr}}$ gives a δE_{corr} adequate to straddle an absorption edge of interest (in this case rhenium L_{III}). We assume point-to-point focusing, $L \ll p, p'$.

The subscript 'corr' refers to the fact that there is a subsequent correlation of the direction of incidence of a ray towards the focus and its (photon) energy. Maier-Leibnitz⁷ and later Thomas⁸ first described the use of a focusing monochromator in this way in a 'modified Laue method' for neutrons.

By choosing appropriate angle of incidence and curvature of the monochromator the band-pass $(\delta\lambda/\lambda)_{\text{corr}}$ of the reflected X-rays can be made to straddle the absorption edge of a spectrum placed at or near the cross-over point. A cylindrical curvature does not exactly reproduce a point focus from a point source but instead a 'caustic'. However, for $L \ll p, p'$ this error is quite negligible². Rays from different points of a finite source cross over at different points so that with such a source there is a finite focus width. However, an accurate correlation between direction and wavelength is maintained for all rays which pass through a sufficiently narrow slit placed at the focus.

Matsushita⁹ has used this property of the bent crystal monochromator for recording photographically the absorption spectrum of the K edge of a copper specimen placed at the cross-over point. The transmitted diverging X-beam produced a streak in which distance and wavelength were correlated.

We now suggest that, if a single crystal is placed near the cross-over point a diffraction pattern is produced in which each spot is drawn out into a streak in which the wavelength varies along the streak (Fig. 1). The theoretical basis for this has been established^{10,11}. Every spot in the pattern then contains a complete energy profile across the absorption edge of one of the atoms in the specimen crystal. This method of producing the whole of all the streaks simultaneously makes much more efficient use of the synchrotron radiation spectrum than the rocking-monochromator streaks method proposed earlier¹² or the more usual approach of measuring whole diffraction data sets at a few discrete wavelengths either with a double crystal monochromator¹ or the current monochromator at the 'Guinier position'¹².

In our preliminary experiments we used the SRS work-station with no modification other than that of placing a narrow adjustable slit at the 'front' end of a large diameter collimator (≈ 1 mm) mounted in the collimator bracket of an oscillation camera¹³. The X-ray beam was therefore effectively unrestricted in vertical height. In the horizontal plane the crystal sample was placed after the cross-over point so that the (now) diverging beam bathed most of the crystal, which also, therefore, accepted the full correlated energy bandwidth available (Fig. 1). The use of the narrow slit (Fig. 1) constrained the energy resolution within the diffraction spot profile to allow satisfactory sampling of the absorption edge fine structure^{2,11}.

We have tested our idea on a single crystal, in the form of a thin plate (0.1×1 mm approximately), of heptahydrido bis-diisopropylphenylphosphine rhenium, $(\text{ReH}_7(\text{P}^i_2\text{Ph})_2)$. This crystal structure has been solved¹⁴ previously from low temperature X-ray data collected on a Nicolet P3/m 4-circle automated diffractometer with graphite monochromatized MoK_α X-ray irradiation.

The space group is $P2_1/n$ with refined cell parameters $a = 11.271 \text{ \AA}$, $b = 13.405 \text{ \AA}$, $c = 17.017 \text{ \AA}$, $\beta = 95.53^\circ$. There is one rhenium atom, 2 phosphorus, 24 carbon and 45 hydrogen atoms in the asymmetric unit. Because the space group is

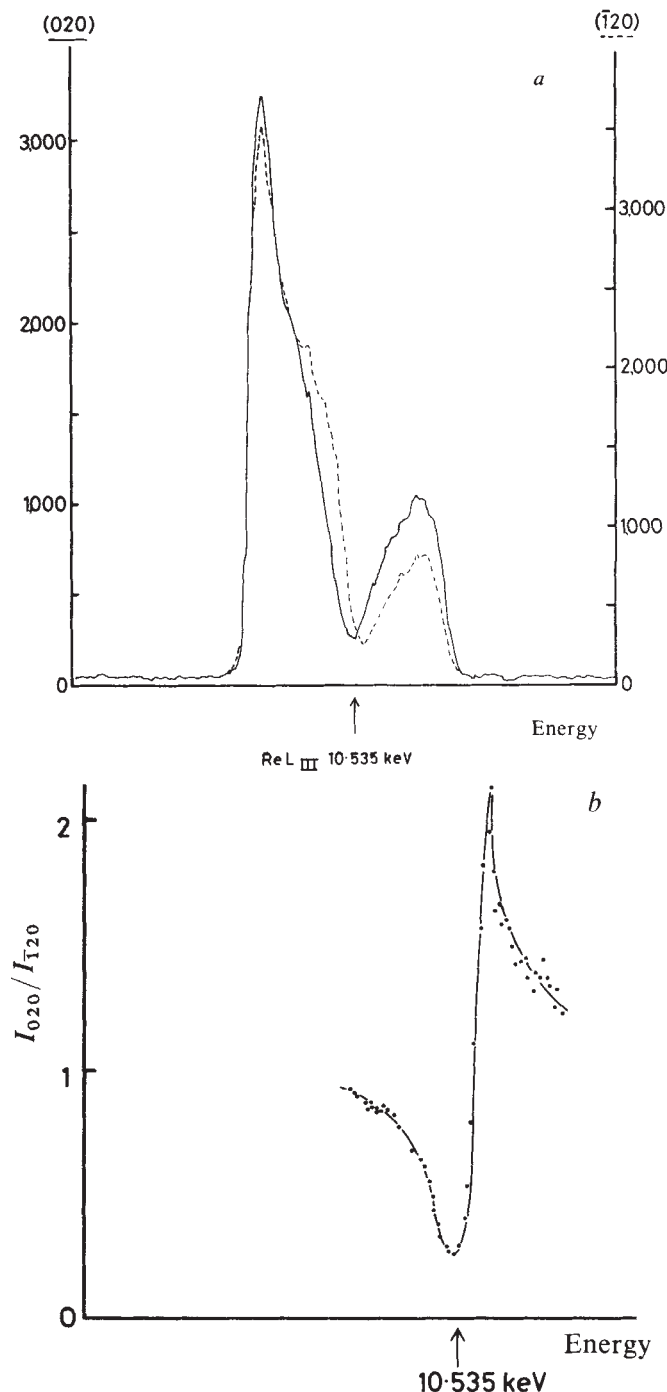


Fig. 2 *a*, Densitometer profiles of the two reflections with Miller indices $\bar{1}20$ and 020 , for the 'on-edge' case. The ordinate scales have been selected to allow exact comparison of reflection widths and the position of the minima in the profiles. The 020 profile is shown as a solid line and $\bar{1}20$ as a broken line. *b*, The divided intensity profiles ($020/\bar{1}20$), after removal of background, showing the effect of the expected dramatic variation due to f'' 'on-edge'.

centrosymmetric, all phases are 0 or π and the variation in f'' has a negligible effect on the reflection intensity. At the SRS we aimed therefore to optimise the value of f' for the rhenium L_{III} edge ($\lambda = 1.772 \text{ \AA}$) across a diffraction spot energy profile with a total bandwidth of 67 eV. The monochromator was a triangular Si(111) monochromator 200 mm long (length in use 156 mm) with a measured oblique cut angle $\alpha = 6.73^\circ \pm 0.05^\circ$ (J.R.H., unpublished data) giving $\delta E_{\text{corr}} = 67 \text{ eV}^{10}$ for $p = 20.9 \text{ m}$, $p' = 2.4 \text{ m}$ and $R = 45 \text{ m}$. We recorded two oscillation photographs, one with the 67 eV centred on the point of inflection of the L_{III} absorption edge of the rhenium in the sample and the other with the 67 eV centred 200 eV below the edge. These two photographs can be compared in ref. 11. The wavelength was calibrated by step scanning the monochromator in 0.01° steps ($\delta E = 9 \text{ eV}$) and measuring the intensity transmitted through a different but otherwise identical single crystal. At the two selected wavelengths about the rhenium edge the sequence of photographs produced was as follows; still 0° , still 85° , oscillation photograph 85–89.95, still 89.95. Exposure time for the 85–89.95° photographs was 5 seconds per degree with 10 oscillations. The crystal to film distance was 54.5 mm corresponding to an outer resolution limit on a flat plate cassette of $1.5 \text{ \AA}$ at this wavelength. The film used was low fog CEA Reflex 25 with three films in a data pack. The SRS was operating at 1.9 GeV with an injection current of $\approx 100 \text{ mA}$ at the time of this experiment.

The diffraction patterns were indexed on the basis of a refined sample orientation (from the stills) taking account of the direction–energy correlation geometry¹⁰ which introduces important changes compared with a conventional X-ray line source¹⁵. Because of the small unit cell, the same reflections are stimulated at both wavelengths though their positions on each film are slightly altered. In the on-edge photograph considerable fine structure is exhibited in the Bragg reflections. For a small heavy atom complex of this sort the absorption profile of the rhenium dominates the different intensities. This causes an apparent uniformity of fine structure between all reflections. The variations in spot-size and inclination for different reflections for this can era geometry is described elsewhere^{10,11}. The on-edge photograph was recorded first; the fine structure recorded is not, therefore, due to a split crystal.

There is a considerable amount of information in these 'three-dimensional' rotat on photographs (x , y , and energy). In this first report of the technique we have only analysed two reflections with the Miller indices 020 and $\bar{1}20$; both reflections are fully recorded. These reflections occur at a closely similar $\sin \theta$ value so that their absorption corrections as a function of λ across the edge are identical to within 1%, even at maximum absorption. However, in the $\bar{1}20$ reflection the rhenium atom happens not to dominate the structure amplitude. The reflection profile of the $\bar{1}20$ is therefore determined by the absorption profile. In the 020 reflection the rhenium does dominate the structure amplitude, hence the expected variation in f' across the edge (as well as absorption) will affect the 020 reflection intensity profile.

The reflections each at the two wavelengths were digitized using a Scandig flat bed densitometer with an aperture (raster) of $10 \text{ }\mu\text{m} \times 100 \text{ }\mu\text{m}$ with a step size of $10 \text{ }\mu\text{m}$. The four central scans were averaged numerically making the sampling raster effectively $10 \text{ }\mu\text{m} \times 400 \text{ }\mu\text{m}$. The scan direction was along the variable energy direction which was here very close to horizontal^{10,11}. Figures 2a and 3a compare the profiles produced in each case. For the on-edge ($\bar{1}20$) profile the change in linear absorption coefficient is 28 mm^{-1} . Figures 2b and 3b show the reflection profile of the 020 divided by the $\bar{1}20$ profile at each wavelength; an average background was removed in each case. The relationship between wavelength and displacement along the streaks is linear within the 67 eV central position¹⁰. Neglecting the end effects the off-edge division shows a linear variation and the on-edge division does indeed reveal the expected dramatic variation of f' at the rhenium L_{III} edge.

Any estimate of f' from profile division should be based on

a least squares refinement including many measured reflections, since small errors in alignment of the profiles may lead to large errors in f' estimates. This work is in progress, but the two reflections quoted here do provide useful evidence of the efficacy of the technique: they give f' estimates for Re which are in good agreement with those measured for Pr and Sm at SSRL¹⁶. From Fig. 2 we obtain estimates of $-27e^-$ at f'_{min} with an f'_{max} of $-1e^-$ and $-17e^-$ at the lower energy limit of our sampling. The energy width of the variation seen in f' from f'_{min} to f'_{max} is reasonable at 11 eV, although the position of the minimum f' value does not quite coincide with the point of

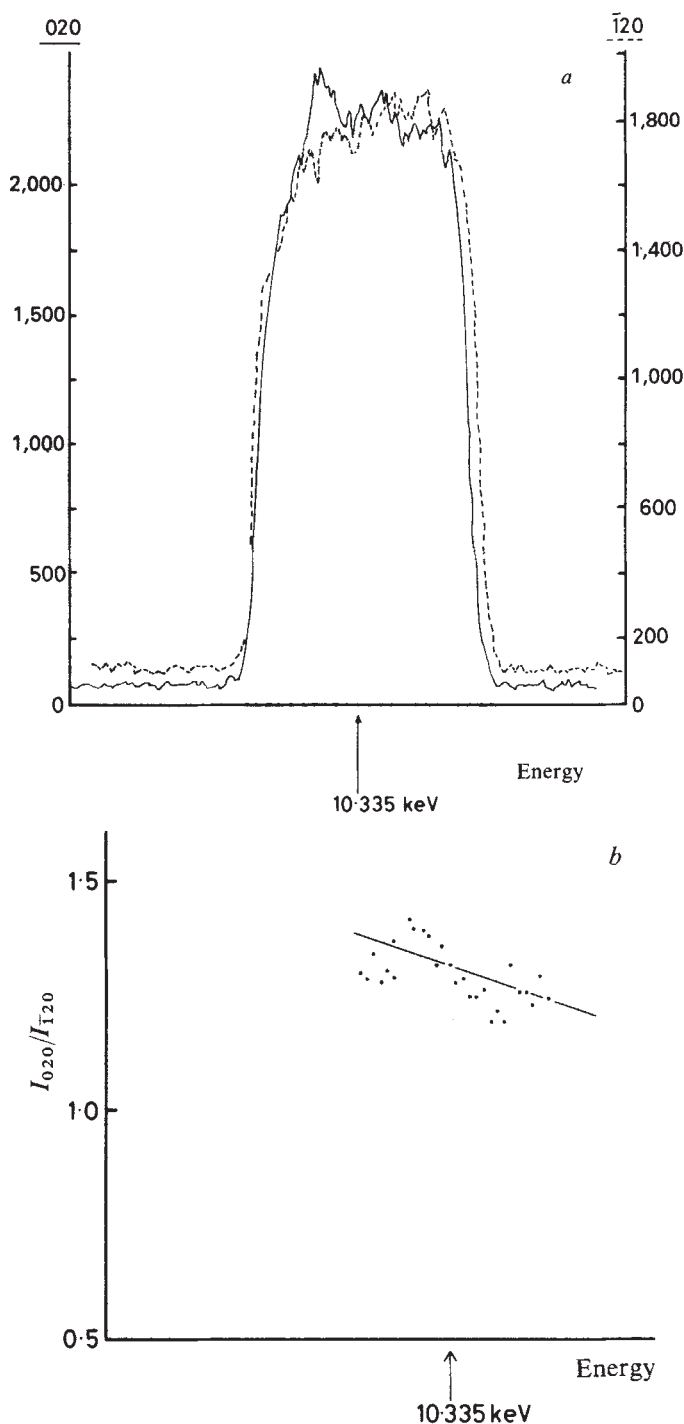


Fig. 3 a, Densitometered profiles of the two reflections $\bar{1}20$ and 020, for the 'off-edge' case. The 020 profile is shown as a solid line and the $\bar{1}20$ as a broken line. b, Divided intensity profiles ($020/\bar{1}20$), after removal of background, showing the effect of a slight linear variation due to f' 'off-edge'.

inflexion of the 120 absorption profile which would have been expected. The energy resolution $\delta E/E$ at any point in the profiles, in this experiment, is of the order of 1×10^{-3} ; this estimate makes allowance for monochromator rocking width, size of slit at focus, thickness of sample, sample mosaic spread, vertical sample size, beam cross-fire angles, correlated wavelength spread and slit-sample-film distances¹¹. There are constraints on these quantities to preserve the energy resolution in a diffraction streak and their effects depend on the location of the spot on the film¹¹. The necessary use of a thin sample also reduced absorption effects, but even so in general absorption corrections will be needed which will require the determination of a complete absorption surface for all energies in the pass-band. These can be determined by recording direct beam transmission streaks as a function of spindle angle ϕ and inclination angle μ of the rotation camera.

We believe that the method has the greatest potential for phasing the reflections from protein crystals containing only one anomalous scatterer per asymmetric unit: in such a case the mass-fraction of the heavy atom is very small (typically $\approx 1\%$); the change in the absorption of the sample at the edge is, therefore, much smaller than with our trial heavy atom complex. Although the heavy-atom contribution to protein structure amplitudes is very small, the atomic scattering factor, f_0 , falls off rapidly with increasing $\sin \theta/\lambda$ whilst the values of f' and f'' of an anomalous scatterer do not; for high-angle reflections for a typical protein crystal the anomalous effect may be as great as 40% of the average structure amplitude (U.W.A., unpublished calculation.) The relative accuracy to which an intensity absorption correction would need to be determined for a protein metal atom derivative would be $\approx 5\%$.

It may be thought that order-to-order resolution for broad diffraction spots may be lost for a dense diffraction pattern from a protein crystal. The change in inclination of the reflection streaks around an oscillation photograph does in fact preserve order-to-order resolution parallel to the horizontal mounting axis outside the blind region (compare for example native protein crystal photographs at the Guinier position and off Guinier position in ref. 10). The overlap perpendicular to the mounting axis is controlled in any case by the size of the rotation angle. However, the use of a small raster step is an inconvenience and it would be necessary to reduce the volume of data produced with a flat-bed densitometer by resorting to an on-line pre-prediction control programme.

We thank MRC and SERC for support, and the SERC Daresbury Laboratory for providing synchrotron experimental beam time and facilities. We also thank Dr J. Spencer of the Department of Inorganic Chemistry, Bristol University for supplying suitable samples, Dr M. Elder of SERC Daresbury Laboratory for providing film densitometry facilities, and Drs S. Ealick of the University of Alabama and G. Taylor of Birkbeck College, London University for helpful discussions.

Australian seismic refraction results, isostasy and altitude anomalies

Peter Wellman

Bureau of Mineral Resources, PO Box 378, Canberra, Australia 2601

Although there have been extensive seismic refraction measurements over areas of continental crust, the causes of variation in crustal thickness, and of variation in mean crustal seismic velocity, are poorly understood. If there is regional ($3^\circ \times 3^\circ$) isostatic equilibrium of the crust, such as suggested by many studies of regional gravity, then crustal thickness variation should be caused largely by variation in surface altitude and crustal density. Seismic refraction results from Australia are used here to confirm that these crustal isostatic effects are a major (40%) cause of variation of crustal thickness even in a continent of little surface relief. However, there is an equally important regional variation in crustal thickness unrelated to crustal isostasy—this is thought to be due to density variation in the lower lithosphere¹.

It is initially assumed that the Earth's crust is in isostatic equilibrium relative to the uppermost mantle and, in addition, that the upper mantle density (σ_m) does not vary horizontally or vertically from where its upper surface under continents is shallow to where it is deep. Consider any crust of mean density σ , surface altitude h , and depth to the base of the crust below sea level H . A standard crust has a surface at sea level, base H_s below sea level, and mean density σ_s . For isostatic equilibrium²

$$H = \frac{H_s(\sigma_m - \sigma_s) + h\sigma}{\sigma_m - \sigma} \quad (1)$$

Average densities of $\sigma_m = 3.3$ tonne m^{-3} and $\sigma_s = 2.9$ tonne m^{-3} can be assumed², so $\sigma/(\sigma_m - \sigma)$ averages ~ 7.25 . In Australia h is generally < 0.6 km, and σ varies from ~ 2.75 to 3.06 tonne m^{-3} , so the effect of h on H can be evaluated with an error of < 2 km.

We define a depth of the base of the crust when corrected for surface altitude as H' ($H' = H - 7.25h$)

$$H' = \frac{H_s(\sigma_m - \sigma_s)}{\sigma_m - \sigma} \quad (2)$$

The Nafe-Drake empirical correlation between rock density and seismic P-wave velocity³ is almost linear in the velocity range 5.5 – 8 $km\ s^{-1}$. Assuming linearity, using velocity subscripts with the same meaning as for density, and inverting equation (2) gives

$$\frac{1}{H'} = \frac{V_m}{H_s(V_m - V_s)} - \frac{V}{H_s(V_m - V_s)} \quad (3)$$

so there should be linear correlation between $1/H'$ and mean crustal velocity V .

Table 1 and Fig. 1 give those crustal seismic refraction results for Australian continental crust where both mean P-wave velocity of the crust and crustal thickness have been determined. Results from regions A, B, C and D are considered less accurate because only travel-times of critically refracted waves were used in the interpretation. Other results were derived using all available crustal and upper mantle seismic P phases. The locations given are those of the crustal thickness determinations, not the shot points.

A least-squares straight-line fit of mean crustal velocity with the inverse of the corrected depth of the base of the crust below sea level for all data is

$$1/H' = (0.123 \pm 0.019) - (0.015 \pm 0.003)V \quad (4)$$

where the uncertainty is the coefficient of variation. The correlation coefficient ($r = 0.59$) is statistically significant at the 95%

Received 30 March; accepted 20 June 1982.

- Phillips, J. C. & Hodgson, K. O. in *Synchrotron Radiation Research* (eds Winick, H. & Doniach, S.) (Plenum, New York, 1980).
- Helliwell, J. R. *et al. J. Phys. E* (in the press).
- Underwood, J. H. & Turner, D. *Soc. Photoopt. Inst. Eng.* **106**, 125 (1977).
- Lemmonier, M., Fournier, R., Rousseau, F. & Kahn, R. *Nucl. Inst. Meth.* **152**, 173–177 (1978).
- Hendrix, J., Koch, M. J. & Bordsas, J. *J. appl. Crystallogr.* **12**, 467–472 (1979).
- Rosenbaum, G. & Holmes, K. C. in *Synchrotron Radiation Research* (eds Winick, H. & Doniach, S.) (Plenum, New York, 1980).
- Maier-Leibnitz, H. *Ann. Acad. Sci. Fenn. A VI*, p. 267 (1967).
- Thomas, P. *J. appl. Crystallogr.* **5**, 83–88 (1972).
- Matsushita, T. *Am. Inst. Phys. Conf. Rep.* **64**, 109 (1980).
- Greenhough, T. J. & Helliwell, J. R. *J. appl. Crystallogr.* (in the press).
- Greenhough, T. J., Helliwell, J. R. & Rule, S. A. *J. appl. Cryst.* (submitted).
- Arndt, U. W. *Nucl. Inst. Meth.* **152**, 307–311 (1978).
- Arndt, U. W., Champness, J. N., Phizackerley, R. P. & Wonacott, A. J. *J. appl. Crystallogr.* **6**, 457–463 (1973).
- Howard, J. A. K., Mead, K. A. & Spencer, J. L. *Acta crystallogr. B* (submitted).
- Greenhough, T. J. & Helliwell, J. R. *J. appl. Crystallogr.* (in the press).
- Templeton, L. K., Templeton, D. H. & Phizackerley, R. P. *J. Am. chem. Soc.* **102**, 1185–1186 (1980).

confidence level. Errors in V are thought to be relatively small, and in equation (4) are assumed to be zero. If errors are partitioned equally between V and $1/H'$ this infers very large errors in the velocity measurements. For example, Pilbara mean crustal velocity would have a systematic error of 0.3 km s^{-1} ; this is unacceptably high either as an error in the mean velocity, or as an error in the density-velocity relationship.

The following calculations suggest that equation (4) is consistent with the assumptions. For these calculations I have adopted the mean values of quantities in Table 1, but with coefficients of variation increased by a factor of 3 to allow for sampling bias

$$V_m = 8.07 \pm 0.08 \text{ km s}^{-1}, \quad V_s = 6.43 \pm 0.11 \text{ km s}^{-1}, \\ H_s = 37.8 \pm 3.6 \text{ km}$$

When H' is infinite, then equation (4) gives an estimate of the mean uppermost-mantle velocity (P_n) of $8.3 \pm 2 \text{ km s}^{-1}$; this is

consistent with the mean of the measured uppermost-mantle velocities. (Note that mantle velocities were not used in equation (4).) The slope of equation (4) is consistent within experimental error with the value of 0.016 ± 0.002 found by evaluating $1/(H_s(V_m - V_s))$ using the means of the measured values given above. It is also consistent within experimental error with the value of 0.020 ± 0.003 found by evaluating equation (3) by using the Nafe-Drake empirical relationship between crustal density and crustal velocity, together with Woollard's² preferred range of values H_s between 33.2 and 34.5 km, and $\sigma_m - \sigma_s$ between 0.383 and 0.415 tonnes m^{-3} .

The least-squares correlation of mantle P_n velocity with depth of the base of the crust below sea level (H) is $V_m = (7.76 \pm 0.14) + (0.0078 \pm 0.0033)H$. Hence there is a systematic change in V_m from 7.99 km s^{-1} at 30 m depth to 8.15 km s^{-1} at 50 km depth; this is equivalent to an increase in mantle density of $0.04 \text{ tonnes m}^{-3}$. This small change would have an effect of less than $<1 \text{ km}$ on H' in equation (4). Regions B, C,

Table 1 Australian crustal seismic refraction results

	Lat. (°S)	Long. (°E)	Altitude h (km)	Mean crustal velocity V (km s^{-1})	Crustal thickness $H + h$ (km)	Mantle velocity (km s^{-1})	Ref.
A Gawler craton							
	31	133	0.04	6.4	42	8.00	14
	30.5	130.5	0.16	6.3	37	8.16	14
B North-east orogens							
	20.2	143.3	0.38	6.25	38.4	7.79	15
	18.2	145.1	0.53	6.31	44.6	7.79	15
	16.4	143.1	0.12	6.31	49.3	8.04	15
	16.5	141.8	0.05	6.06	35.6	8.04	15
	11.6	142.5	0.01	6.34	36.4	7.95	15
	13.8	142.1	0.12	6.32	36.6	7.95	15
C Southern Yilgarn block							
	31.4	120.2	0.41	6.50	35.5	8.34	16
	31.7	117.0	0.28	7.05	43.0	8.34	16
	35.0	118.0	0.00	6.43	34.0	8.06	16
D Central Tasman fold belt							
	22.8	148.7	0.16	5.29	35.0	8.1	17
	24.0	149.4	0.15	6.31	37.0	8.1	17
E Adelaide geosyncline							
	31.0	138.5	0.38	6.19	38.5	7.97	18
	33.2	138.0	0.10	6.30	38.5	7.97	18
	32.5	138.7	0.43	6.25	34.0	7.97	18
F Southern Lachlan geosyncline							
	35.9	148.4	1.37	6.61	50.5	8.02	19
	35.3	149.5	0.78	6.70	50.5	8.02	19
	35.8	147.4	0.43	6.65	42	8.05	20
	35.8	147.7	0.54	6.62	44.5	8.04	20
	36.7	148.6	0.62	6.65	50	8.03	20
	33.2	150.8	0.26	6.64	43.5	8.05	21
	34.2	150.4	0.31	6.68	41.5	8.12	21
	34.8	149.3	0.67	6.52	43	7.95	21
G Pilbara block							
	21	119.8	0.23	6.32	28	8.34	22
	22.8	120	0.46	6.28	33	8.34	22
	22	117	0.54	6.37	32	7.79	4
	22.8	117.8	0.53	6.43	31	7.79	4
	23.0	119.2	0.67	7.29	30	7.79	4
	20.8	119.2	0.15	6.32	28	8.18	23
	22.4	118.2	0.71	6.29	30	8.18	23
	24	117.9	0.47	6.52	40	8.10	23
	23.8	117.2	0.40	6.62	40	8.10	23
H Northern Yilgarn block							
	25.9	118.9	0.55	6.51	52	8.3	22
	26	118.2	0.46	6.60	55	8.10	23
I SE North Australian craton							
	20.4	139	0.32	6.70	55	8.23	24
	19.4	134.9	0.26	6.67	51	8.15	24
	16.7	135.5	0.20	6.75	52	8.47	25
	16.3	134	0.25	6.72	52	8.47	25
	16.5	137	0.10	6.83	52	8.00	25
J North Northern Australian craton							
	10.6	130.2	-0.04	6.01	31.4	7.91	26
	10.4	130.7	-0.04	6.02	34.2	7.91	26
	4.6	135.5	-0.03	5.91	26.6	7.63	26
	10.5	130.5	-0.03	6.14	34	8.00	27

E, F, Q and J have mantle velocities measured from directions near right angles. Significant upper mantle anisotropy is suggested only by results from the Pilbara Block⁴.

I conclude that the effect on crustal thickness of variations of surface altitude and mean crustal density can be adequately modelled using equation (3). This equation assumes regional isostatic equilibrium relative to the uppermost mantle, assumes that the mantle has constant density down to the level of the deepest crust, and assumes a linear correlation between P-wave velocity and density.

The relative magnitude of the components of crustal thickness variation are evaluated as follows. The total variance from the measured crustal thicknesses listed in Table 1 is 66.0 km^2 . Surface altitude variation and its associated crustal root variation is calculated from the difference in variance of H and H' to be 2% of this. The change in crustal thickness associated with change in mean crustal density is $\sim 38\%$ of the variance ($r^2 = 0.38$ for equation (4)). The remaining variance (that causes deviation from equation (4)) can be split into within-region variance of 11.3 km^2 (17% of total variance), and between-region variance causing the remaining 43%.

The within-region variance is caused both by errors in determination of relative crustal thickness and relative mean crustal velocity within each region, and by crustal strength being sufficiently high in western and central Australia that areas $< 3^\circ \times 3^\circ$ are not in complete isostatic equilibrium⁵. The variance due to departures from isostatic equilibrium is unlikely to be $> 8\%$ of total variance (one-half within-region variance) and so is much smaller than the variance due to the isostatic effects of topography and crustal density changes (40% of total variance). This reinforces the conclusion that isostatic equilibrium relative to the uppermost mantle is largely achieved.

Between-region variance is $\sim 43\%$ of total variance, so its origin is important both in understanding and predicting crustal thickness. It is unlikely that errors in crustal thickness and crustal velocity contribute significantly to within-region or between-region variance, because the within-region variance is small both in the better studied regions (F, G and I) and in regions where independent investigators have made independent measurements (regions I and J). Between-region changes in mean atomic number could, in theory, cause regional changes in the velocity-density relationship^{6,7}. There is no good evidence for significant differences in the chemistry of the surface rocks from the various regions so significant change in mean atomic number seems unlikely. The remaining possible causes of between-region variance—horizontal density variations in the lower lithosphere, and dynamic forces on the base of the litho-

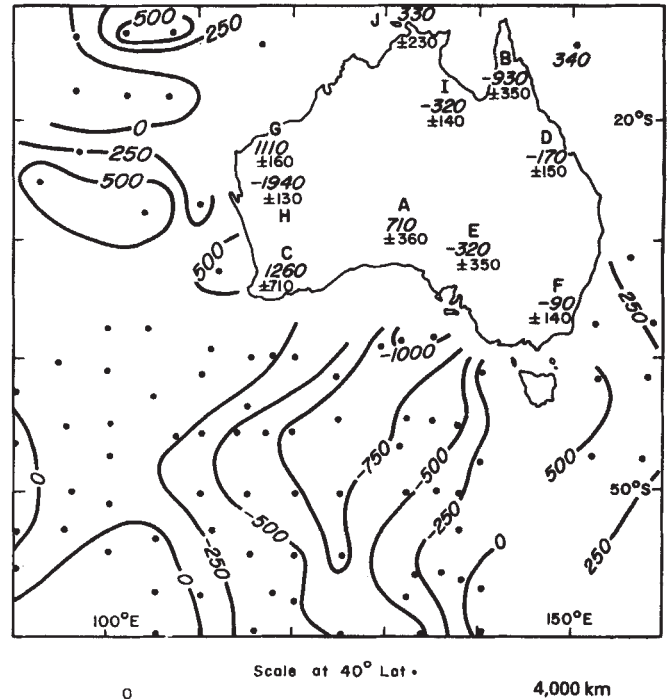


Fig. 2 Mean height anomalies in metres. Shallower areas have positive mean height anomalies. In oceanic areas dots show data points; contour interval, 250 m.

sphere—are considered below by comparing mantle altitude anomalies under continents with those under oceans.

The between-region differences in H' residual are converted to equivalent differences in the upper level of mantle overlain by oceanic crust of assumed constant thickness and by seawater of density σ_w . A difference of p in the thickness of sea-level crust is equivalent to a difference of $p(\sigma_m - \sigma_w)/(\sigma_m - \sigma_w)$, or about $p/5.68$, in the upper level of oceanic mantle. For each region of Fig. 1, a mean altitude anomaly and its coefficient of variation has been calculated using H' residuals divided by 5.68.

In oceanic areas, altitude anomalies are calculated by subtracting from $1^\circ \times 1^\circ$ mean water depths, the expected water depth for that age crust, and correcting for the isostatic effect of sediment thickness¹. The altitude anomalies used are the means of $5^\circ \times 5^\circ$ areas. Altitude anomalies derived for continental areas are assumed to have the same mean (zero) as those derived for areas of oceanic crust.

Figure 2 shows altitude anomalies derived from seismic refraction data in continental Australia, together with published altitude anomalies for the Indian and Southern Oceans¹—these have been recontoured west of Australia. East of Australia in the Tasman and Coral Seas new rough estimates of altitude anomalies are given.

Altitude anomalies do not show a systematic trend across Australia, nor do they show a correlation with seismic P-wave residuals⁸⁻¹⁰, heat flow variation¹¹, or long-wavelength gravity anomalies¹.

Continental and oceanic values do not show an overall pattern, as there seems to be a break between continental and oceanic values. The lack of systematic variation across the region indicates that altitude anomalies are unlikely to be due to dynamic asthenospheric forces on the base of the lithosphere as these would be continuous across lithosphere boundaries. The interpretation that altitude anomalies originate from horizontal density variation in the lower lithosphere¹ is therefore supported. Relative to oceanic altitude anomalies, those from the continent have a larger range and shorter wavelength; this supports models where the continental lithosphere has had a relatively long and diverse history¹². An analysis of seismic

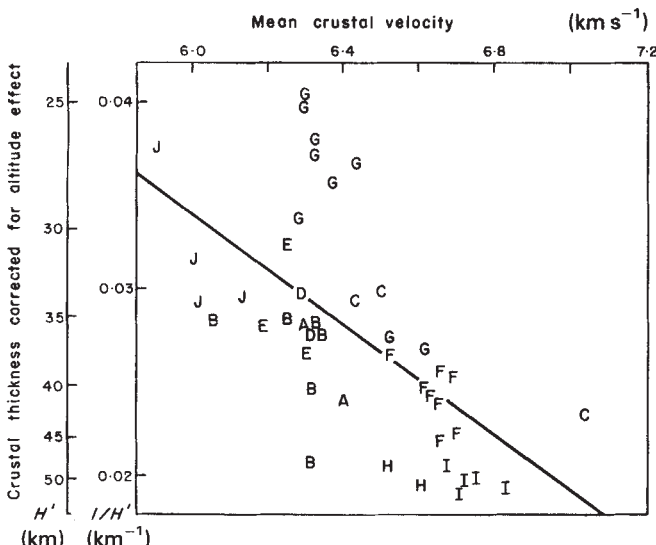


Fig. 1 Relation of inverse of corrected crustal thickness to mean crustal velocity. The letters refer to the regions of Table 1 and Fig. 2. The line gives the least-squares correlation.

refraction data from all continents would define the constants of equation (3) more precisely, and allow altitude anomalies to be mapped over the whole Earth.

I interpret variation in crustal thickness as being mainly due to two effects. (1) Because surface altitude and mean crustal density vary relatively rapidly on continents, the crust appears to be in isostatic equilibrium with an upper mantle of uniform density, so variation in surface altitude and mean crustal density

show a strong correlation with variation in mean crustal thickness. (2) Isostatic equilibrium is actually between the whole lithosphere and the asthenosphere¹³, so, to keep equal pressure on the asthenosphere and keep the continental crust near sea level, any regional variation in lower lithosphere density causes regional variation in continental crustal thickness.

This work is published with the permission of the Director, Bureau of Mineral Resources, Canberra.

Received 17 March; accepted 29 June 1982.

1. Cochran, J. R. & Talwani, M. *Geophys. J. R. astr. Soc.* **50**, 495–522 (1977).
2. Woollard, G. P. *Phys. Earth planet. Inter.* **3**, 484–498 (1970).
3. Talwani, M., Sutton, G. H. & Worzel, J. O. *J. geophys. Res.* **64**, 1545–1555 (1959).
4. Drummond, B. J., Smith, R. E. & Horwitz, R. C. *Geol. Soc. Aust. Spec. Publ.* **7**, 33–41 (1981).
5. Wellman, P. *BMJ J. Aust. J. Geol. Geophys.* **1**, 297–302 (1976); **3**, 153–162 (1978); **4**, 373–382 (1979).
6. Manghnani, M. H., Ramanantoandro, R. & Clark, S. P. *J. geophys. Res.* **79**, 4527–4546 (1974).
7. Drummond, B. J. & Shelley, H. M. *BMJ J. Aust. J. Geol. Geophys.* **6**, 137–143 (1981).
8. Cleary, J. *Bull. seis. Soc. Am.* **57**, 773–781 (1967).
9. Everingham, I. B. *Bur. Min. Resour. Aust. Rec.* 1969/22 (1969).
10. Cleary, J. R., Simpson, D. W. & Muirhead, K. J. *Nature* **236**, 111–112 (1972).
11. Cull, J. P. *BMJ J. Aust. J. Geol. Geophys.* **7**, 11–21 (1982).
12. Jordan, T. H. *Nature* **274**, 554–548 (1978).

13. Dooley, J. C. *BMJ J. Aust. J. Geol. Geophys.* **2**, 1–5 (1977).
14. Cleary, J. R. *Tectonophysics* **20**, 241–248 (1973).
15. Finlayson, D. M. *J. geol. Soc. Aust.* **15**, 35–50 (1968).
16. Mathur, S. P. *Tectonophysics* **24**, 151–182 (1974).
17. Collins, C. D. N. *BMJ J. Aust. J. Geol. Geophys.* **3**, 203–209 (1978).
18. Shackelford, P. R. J. & Sutton, D. J. *J. geol. Soc. Aust.* **28**, 491–500 (1981).
19. Finlayson, D. M., Collins, C. D. N. & Denham, D. *Phys. Earth planet. Inter.* **21**, 321–342 (1980).
20. Finlayson, D. M., Prodehl, C. & Collins, C. D. N. *BMJ J. Aust. J. Geol. Geophys.* **4**, 243–252 (1979).
21. Finlayson, D. M. & McCracken, H. M. *J. geol. Soc. Aust.* **28**, 177–190 (1981).
22. Drummond, B. J. *BMJ J. Aust. J. Geol. Geophys.* **4**, 171–180 (1979).
23. Drummond, B. J. *BMJ J. Aust. J. Geol. Geophys.* **6**, 123–135 (1981).
24. Finlayson, D. M. *J. geophys. Res.* (in the press).
25. Collins, C. D. N. *BMJ J. Aust. J. Geol. Geophys.* (in the press).
26. Jacobson, R. S., Shor, G. G., Kieckhefer, R. M. & Purdy, G. M. *Am. Ass. petrol. Geol. Mem.* **29**, 209–222 (1979).
27. Rynn, J. M. W. & Reid, I. D. *J. geol. Soc. Aust.* (in the press).

Vertical distribution and isotopic fractionation of living planktonic foraminifera from the Panama Basin

Richard G. Fairbanks*, Marc Sverdløve*,
Rosemary Free*, Peter H. Wiebe†
& Allan W. H. Be*

* Lamont-Doherty Geological Observatory of Columbia University, Palisades, New York 10964, USA

† Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA

A model^{1,2}, which explains the vertical and seasonal distribution of planktonic foraminifera was developed using data from the western North Atlantic and is comprised of three tenets: (1) The conditions in the photic zone (upper 80–100 m) exclusively dictate planktonic foraminifera species composition and abundance throughout the water column. (2) Species are vertically stratified within the photic zone according to their temperature preference. Depth distribution according to temperature preference has been proposed by several authors³ based on isotope analyses of sediments. Due to seasonal temperature variations, seasonal variations in the $\delta^{18}\text{O}$ composition of seawater, and species-specific isotope fractionation, rigorous proof of this hypothesis cannot be provided from sediment analyses alone. (3) The chlorophyll maximum is a major food source zone, which is preferentially exploited by foraminifera. Our data from the Panama Basin, which are presented here, are confirmation of this model. The Panama Basin in the eastern equatorial Pacific, which has a shallow and steep thermocline, was sampled for living planktonic foraminifera. A multiple opening-closing net and environmental sensing system (MOCNESS)⁴ was used to collect 13 sets of eight vertically stratified oblique samples. We report here the quantitative distribution and stable isotopic composition of planktonic foraminifera from four MOCNESS tows taken over 5 days, during August 1979 (RV *Knorr* cruise 73) at 5°20' N, 8°50' W. Each tow consisted of eight samples which integrated ~12.5-m intervals from 0 to 100 m for MOC 126, 25-m intervals from 0 to 200 m for MOC 127, 100–150-m intervals from 0 to 1,000 m on MOC 128 and 250-m intervals from 0 to 2,000 m on MOC 131.

Oxygen and carbon isotope analyses were made on planktonic foraminifera from the MOCNESS tows and the surrounding seawater. Isotope determinations were made on a triple

collector mass spectrometer (Micromass 903). The carbonate samples were vacuum roasted at 375 °C for 1 h and reacted with 100% phosphoric acid at 50 °C. The precision for carbonate is $\pm 0.05\%$ for oxygen and carbon. The $\delta^{18}\text{O}$ composition of seawater was determined with a rapid equilibration-extraction method described in ref. 5. The $\delta^{13}\text{C}$ composition of total dissolved CO_2 was predicted using the PO_4 - $\delta^{13}\text{C}$ relationship given in ref. 6 and PO_4 data for a station occupied in between MOCNESS tows. The 0–100-m series provided an ideal test of the previous model^{1,2} for several reasons. (1) Eight MOCNESS oblique tows at 12.5-m intervals (MOC 126) provide high resolution sampling. (2) There is a 15 °C temperature gradient in the upper 100 m, encompassing one of the steepest thermoclines (0.5°C m^{-1}) in the world's oceans which occurs between 25 and 37.5 m. (3) The region is extremely productive, and foraminifera are abundant. (4) A sharp chlorophyll maximum is well developed in the thermocline⁷. (5) The water column was sampled for measurement of $\delta^{18}\text{O}$ of seawater.

Twenty-three species of planktonic foraminifera were identified in the Panama Basin tows. In the 0–100-m tow, MOC 126 (Fig. 1a), 16 species averaged more than 10 per 1,000 m³. The dominant species, *Globoquadrina dutertrei*, showed a pronounced abundance peak between 25 and 50 m, precisely the interval corresponding to the steep thermocline, highest primary productivity and the associated deep chlorophyll maximum (DCM) (Fig. 1a)⁷. Nine out of 16 species or morphotypes showed pronounced maxima in the steep thermocline. Note that the abundance axis is logarithmic. *Globigerinoides sacculifer* and *Globigerinoides conglobatus* were numerically dominant at the chlorophyll maximum, but they showed comparable abundances in the surface mixed layer. *Globigerinoides ruber*, *Globigerinita glutinata*, *G. sacculifer* (without final sac chamber), and *G. conglobatus* were very abundant in the less productive surface mixed layer. Two species which secrete morphologically distinct terminal chambers, *Orbulina universa* and *G. sacculifer*, showed sharp maximum concentrations of these morphologies in the chlorophyll maximum zone between 25 and 37 m (Fig. 1a). *Globigerina calida* and *Globorotalia theyeri* had maxima below the steep thermocline. Of particular significance is the observation that the *G. theyeri* population descended from a maximum at 100 m on 4 August, MOC 127 (Fig. 1b) to between 250 and 375 m on 5 August, MOC 128 (Fig. 1c) to a maximum between 500 and 750 m on 8 August, MOC 131 (Fig. 1d). This averaged to a 170 m per day descent rate.

Oxygen isotope measurements of seawater (Fig. 2) and continuous temperature and conductivity measurements during the MOCNESS tows permit the construction of an accurate calcite

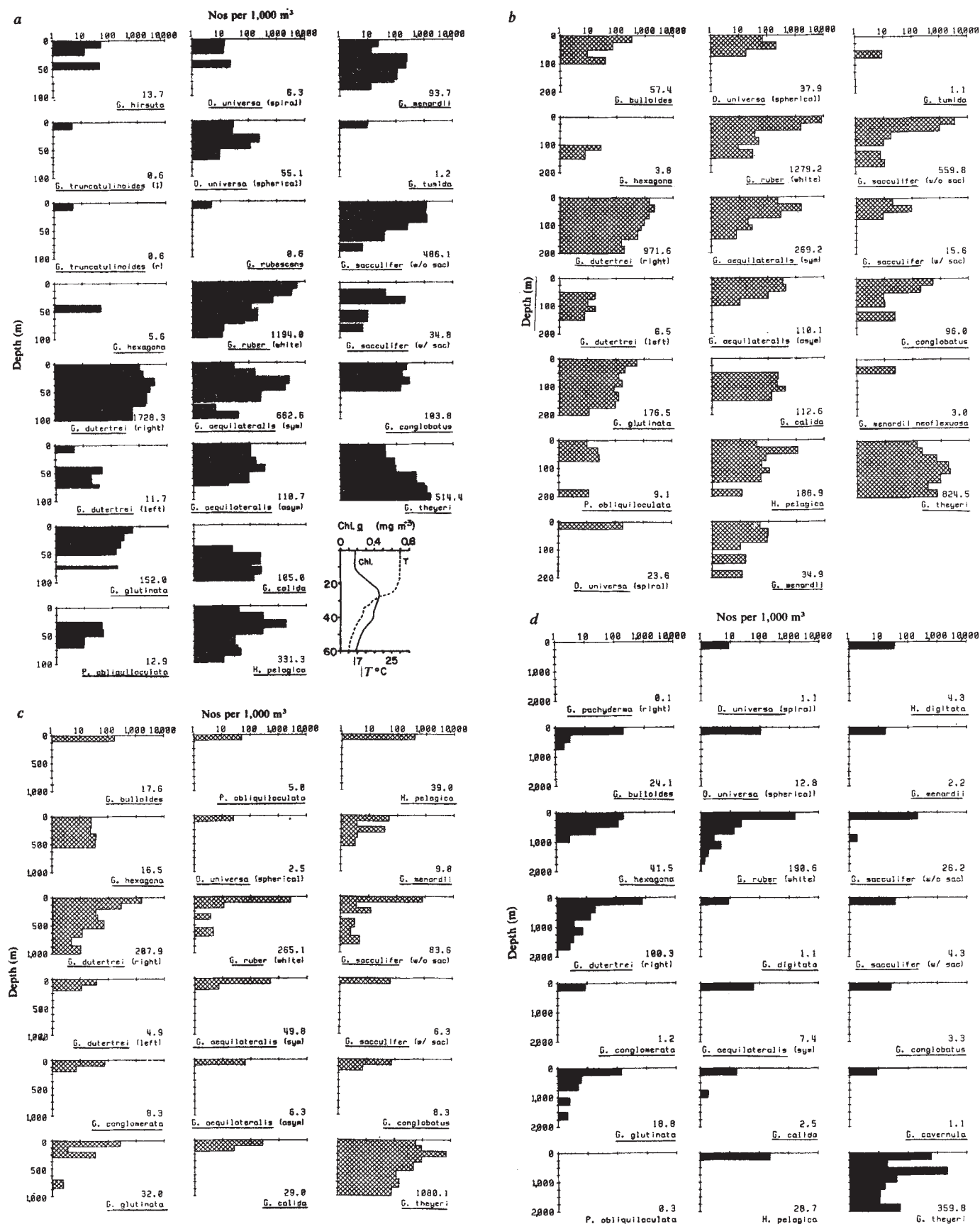


Fig. 1 *a*, Vertical profiles of living planktonic foraminifera caught in MOCNESS 126 oblique hauls in the Panama Basin between 0 and 100 m (3 August 1979; night). The number in the lower right corner of each histogram represents the average number per 1,000 m³ over the interval towed, in this case 0–100 m. Temperature and chlorophyll profiles for this station are in the lower right corner. Chlorophyll data from Marra *et al.*⁷. *b*, Vertical profiles of living planktonic foraminifera caught in MOCNESS 127 oblique hauls in the Panama Basin between 0 and 200 m (4 August 1979; day). *c*, Vertical profiles of living planktonic foraminifera caught in MOCNESS 128 oblique hauls in the Panama Basin between 0 and 1,000 m (5 August 1979; day). *d*, Vertical profiles of living planktonic foraminifera caught in MOCNESS 131 oblique hauls in the Panama Basin between 0 and 2,000 m (8 August 1979; night).

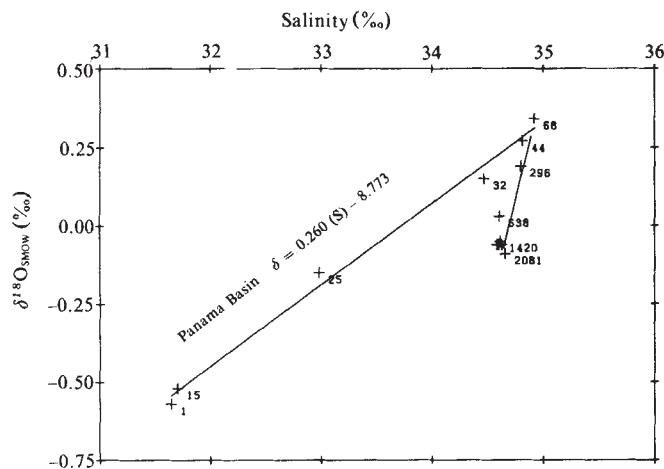


Fig. 2 Salinity versus $\delta^{18}\text{O}_{\text{SMOW}}$ for the Panama Basin. The numbers next to the symbols indicate water depth. The data indicate binary mixing between surface water and intermediate water, and deep water and intermediate water.

equilibrium profile (Fig. 3a) using the mollusc equation of Epstein *et al.*⁸. Oxygen isotope measurements of several different size fractions of nine foraminifera species or morphotypes show that shell growth is restricted to the upper 100 m, even though many species are found in abundance below 100 m. The oxygen isotope data indicate that specimens found below 100 m were simply descending through the water column while adding little or no additional calcite (Fig. 3a). It is also evident from the $\delta^{18}\text{O}$ versus depth plots (Fig. 3a) that most species generally grow within narrower depth intervals within the photic zone than their overall distribution would imply. For example, *G. ruber* is found in decreasing numbers between the surface and 100 m (Fig. 1a). Its $\delta^{18}\text{O}$ composition indicates that it is growing typically 0.5‰ lighter than equilibrium^{1,2,9} in the surface mixed layer. Between 25 and 37.5 m *G. ruber* is adding some calcite, but below this level the data-point at 60 m indicates vertical descent from the mixed-layer.

G. dutertrei calcifies in oxygen isotope equilibrium² and the isotope data indicate growth in the thermocline, where *G. dutertrei* are found in maximum abundance (Fig. 1a). Between 37.5 and 100 m, *G. dutertrei* becomes only slightly enriched in $\delta^{18}\text{O}$, and are far from equilibrium at these depths. This is evidence that they are descending from the steep thermocline (Fig. 3a). The slight $\delta^{18}\text{O}$ enrichment is probably due to shell thickening^{2,3}. Scanning electron microscope observations are needed to confirm this hypothesis.

G. theyeri and *G. calida* appear to be in oxygen isotopic equilibrium between 50 and 100 m, where they are dominant in MOC 126 and MOC 127, but their isotopic values do not change when they are collected below the photic zone (approximately upper 100 m) even though they may be numerically dominant below 100 m. Again, vertical descent is obvious and, in these two examples, no additional calcite was added during descent, otherwise we would observe a shift towards the equilibrium line. Thus we feel confident that this sampling strategy adequately captures all adult stages of shell calcification and allows us to examine the entire calcification process.

There is a 2% decrease in the $\delta^{13}\text{C}$ of total dissolved CO_2 between the surface and 100 m at the site of MOC 126 (Fig. 3b). There is some evidence that the $\delta^{13}\text{C}$ compositions of *G. dutertrei* and *Globigerinella aequilateralis* are affected by this gradient. *G. dutertrei* decreases more than 1% $\delta^{13}\text{C}$ between the surface and 100 m (Fig. 3b). The largest drop in $\delta^{13}\text{C}$ for *G. dutertrei* is coincident with the 1.5% decrease in $\delta^{13}\text{C}$ of total dissolved CO_2 in the steep thermocline. *G. aequilateralis* show a less significant shift in $\delta^{13}\text{C}$, but this species also parallels the seawater trend. The carbon isotope equilibrium line for inorganic calcite would fall to the right of the data, but parallel to the $\delta^{13}\text{C}$ of total dissolved CO_2 line (see ref. 9 for a discussion of calcite isotope equilibrium).

Note that *G. dutertrei* and *G. theyeri*, which calcify in oxygen isotope equilibrium, have the same carbon isotopic composition as ambient $\delta^{13}\text{C}$ of total dissolved CO_2 (Fig. 3b). As noted above, oxygen isotope measurements of *G. dutertrei* indicate equilibrium precipitation at $\sim 30\text{ m}$ (Fig. 3b). Below 30 m *G. dutertrei* shows oxygen isotope enrichment, but the data are

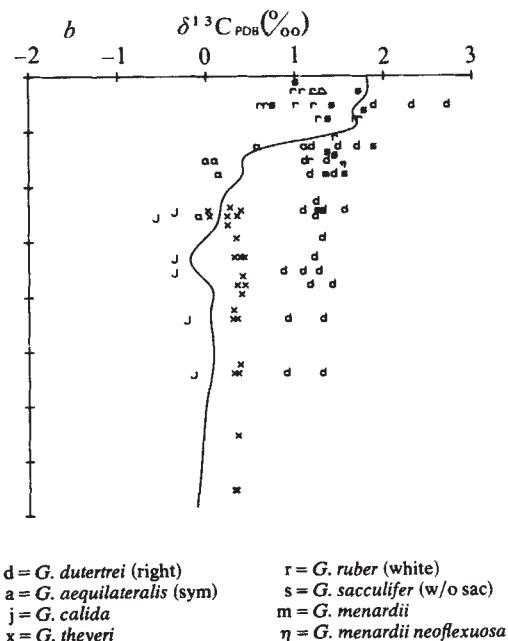
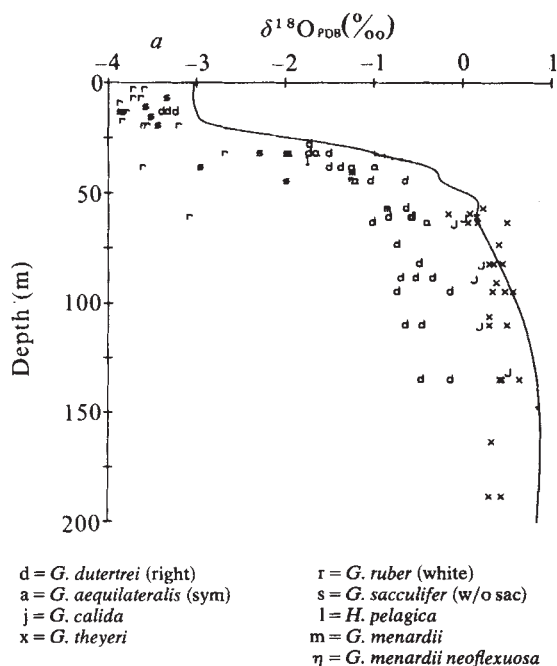


Fig. 3 *a*, Oxygen isotope composition of various size fractions of planktonic foraminifera as a function of depth in MOC 126, 0–100 m and MOC 127, 0–200 m. The solid line is the calculated equilibrium CaCO_3 - $\delta^{18}\text{O}$ variation. All specimens plotted were protoplasm full. Corresponds to histograms in Fig. 1*a, b*. Carbon isotope composition of various size fractions of planktonic foraminifera as a function of depth in MOC 126, 0–100 m and MOC 127, 0–200 m. The solid line is the $\delta^{13}\text{C}$ of total dissolved CO_2 .

significantly lighter than equilibrium, indicating vertical descent from 30 m. Similarly *G. dutertrei* calcite has the same $\delta^{13}\text{C}$ as the $\delta^{13}\text{C}$ of total CO_2 at 30 m and shows departure from the $\delta^{13}\text{C}$ of total CO_2 curve below 30 m, which is in agreement with the $\delta^{18}\text{O}$ data. *G. theyeri* is in oxygen isotope equilibrium between 50 and 75 m and also has the same $\delta^{13}\text{C}$ of carbonate as ambient $\delta^{13}\text{C}$ of total dissolved CO_2 . *G. theyeri* differs from *G. dutertrei* by 1‰. The algal symbiont bearing species (such as *G. ruber*, *G. sacculifer*, *G. aequilateralis*) are generally depleted with respect to ^{13}C when compared with the ambient $\delta^{13}\text{C}$ of total dissolved CO_2 .

We thank Greg Kolibas for stable isotope measurements, Ted Baker for computer programming and J. K. Bishop for providing PO_4 data. Contribution no. 3358 from Lamont-Doherty Geological Observatory of Columbia University and contribution no. 5139 from the Woods Hole Oceanographic Institution. This work was supported by NSF grant OCE 7725976 (R.G.F.) and ONR grants N00014-750210 and N00014-800098.

Received 25 March; accepted 24 May 1982.

1. Fairbanks, R. G. & Wiebe, P. H. *Science* **209**, 1524 (1980).
2. Fairbanks, R. G., Wiebe, P. H. & Bé, A. W. *Science* **207**, 61 (1980).
3. Emiliani, C. *Am. J. Sci.* **252**, 149 (1954).
4. Wiebe, P. H., Burt, K. H., Boyd, S. H. & Morton, A. W. *J. mar. Res.* **34**, 313 (1976).
5. Fairbanks, R. G. *J. geophys. Res.* **87**, 5796 (1982).
6. Broecker, W. S. *Prog. Oceanogr.* **11**, 151 (1982).
7. Marra, J., Wiebe, P. H., Stepien, J. C. & Bishop, J. K. (in preparation).
8. Epstein, S., Bucksbaum, R., Lowenstam, H. A. & Urey, H. C. *Bull. geol. Soc. Am.* **64**, 1315 (1953).
9. Williams, D. F., Bé, A. W. H. & Fairbanks, R. G. *Science* **206**, 447 (1979).
10. Turner, J. V. (in preparation).

Trans-sexually grafted antennae influence development of sexually dimorphic neurones in moth brain

Anne M. Schneiderman*†, Steven G. Matsumoto* & John G. Hildebrand†

* Department of Neurobiology, Harvard Medical School, Boston, Massachusetts 02115, USA

† Department of Biological Sciences, Columbia University, New York, New York 10027, USA

Observations in insects reveal sexual differences in the central nervous system (CNS), associated with sexually dimorphic patterns of reproductive behaviour¹⁻⁹. For example, in certain species of moths, including the sphinx moth *Manduca sexta*, only males fly towards a sexually receptive female or towards a source of the female sex pheromone^{10,11}. In *Manduca*, specialized olfactory receptor cells found only on male antennae^{12,13} respond sensitively and selectively to the female sex pheromone (unpublished experiments with K.-E. Kaissling and R. J. O'Connell). Their axons project into the macroglomerular complex (MGC), which is characteristic of male, but not female, antennal lobes (ALs; Fig. 1b, d)^{1-3,5,8,9,14}. These afferents to the MGC presumably synapse with male-specific AL neurones^{8,15} to begin the processing of phenomonal information. We have now devised a surgical procedure for producing antennal gynandromorphs of *Manduca* in which one of the two ALs receives sensory innervation from an antenna formed by a transplanted imaginal disk of the opposite sex. We report here that in these gynandromorphs, the physiological and morphological properties of certain AL neurones are influenced by the gender of the antennal sensory axons contacting them. In particular, neurones resembling the male-specific AL neurones appear in female ALs innervated by sensory axons from a grafted male antenna.

We produced 329 gynandromorphic larvae (Table 1, group 1), of which 78% metamorphosed and eclosed as adult moths.

In 49% of these moths, the transplanted imaginal disk had everted and formed an antenna. All transplanted antennae exhibited structure characteristic of the donor's sex; none appeared to be modified by transplantation into a host of the opposite sex.

In 59% of the animals with everted grafts, the AL on the operated side was innervated by the nerve from the transplanted antenna. In these animals, the ingrowing sensory axons found their targets, the AL neurones, even though the pupal antennal nerve, comprising axons believed to provide contact guidance for antennal afferents^{13,16}, had been severed during surgery. We presume that the axons followed stumps of the 'guide fibres' still associated with the AL and suspect that the growing axons failed to do so in the remaining animals (41% of those with everted grafts). In the latter, the antennal nerve either ended abruptly as a mass of fibres near the base of the antenna or grew into the lateral protocerebrum, proboscis, or sub-oesophageal ganglion. In these animals, the operated AL was stunted, as in the deafferented AL in a deantennated moth^{17,18}. No innervation of the contralateral AL by the graft was ever observed.

We examined histological sections from the brains of unoperated moths and antennal gynandromorphs (Fig. 1; Table 1,

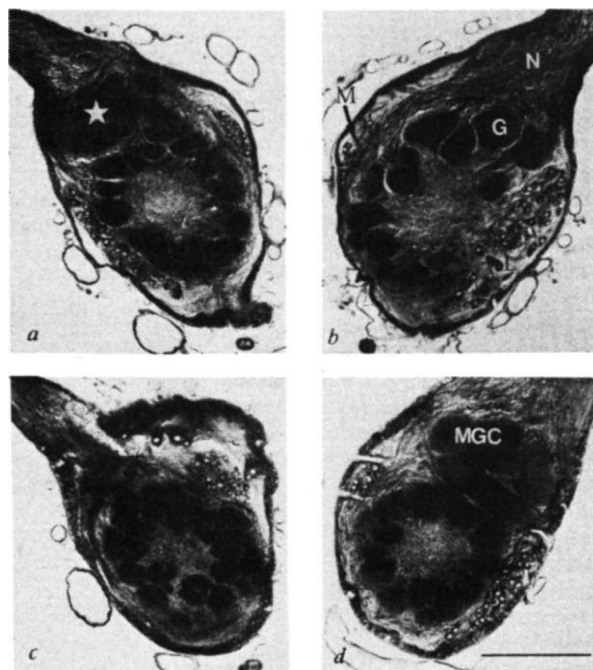


Fig. 1 Frontal sections through the operated (gynandromorphic) and control (normal) ALs of female (a, b) and male (c, d) moths, prepared from paraffin-embedded preparations and stained with Cresylecht violet and Luxol fast blue¹⁸. To produce antennal gynandromorphs of *Manduca*, antennal disks are transplanted unilaterally on the first or second day of the fifth (last) larval instar. Larvae are restrained in aluminium tubes and anaesthetized by chilling in a bath of ice-cold water. The antennal disk¹⁴, which lies within the head capsule near the base of the larval antenna, is excised and replaced with the corresponding disk from a donor of the opposite sex. A mixture of phenylthiourea, streptomycin sulphate and penicillin (2:1:1)¹⁹ is placed in the wound, which is then sealed by melted myristic acid (W. Harris, personal communication). In b, the antennal nerve (N) enters the normal female AL. The glomeruli (G) are arrayed around a central region of coarser neuropil. The cell bodies of the AL neurones are gathered in three cortical groups: a large lateral cluster (L), a smaller medial cluster (M) and a small group anterior to this plane of section. The normal male AL (d) characteristically exhibits the MGC, which lies near the entrance of the antennal nerve and is not present in the normal female AL (b). In the gynandromorphic female AL (a), a structure resembling the MGC (☆) lies near the entrance of the antennal nerve from the grafted male antenna. The gynandromorphic male AL (c) lacks a MGC. Scale bar, 200 μm .

* Present address: Department of Biological Sciences, 901 Fairchild, Columbia University, New York, New York 10027, USA.

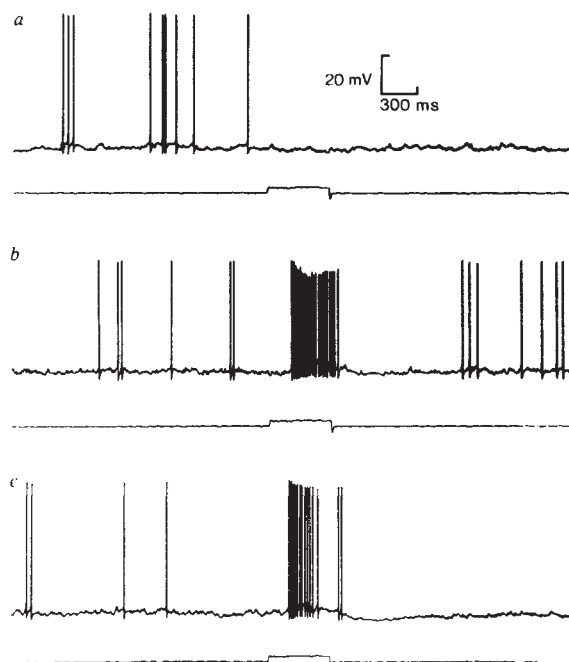


Fig. 2 Intracellular recording from a neurone in the operated AL of a female gynandromorph. Recordings were made with microelectrodes filled with a solution of 4% CoCl_2 in 0.75 M KCl and having impedances of 50–75 M Ω (ref. 8). This recording shows the postsynaptic responses elicited by directing pulses of air (a), *trans*-2-hexenal (b), or bombykal (c) on to the grafted male antenna. Odourants were delivered in carrier pulses (500–600 ms duration) of charcoal-filtered air. This cell was stimulated when the antenna was exposed to *trans*-2-hexenal or bombykal. AL neurones responsive to antennal stimulation with bombykal were not found in unoperated females or in the control AL of gynandromorphic females. Conversely, in gynandromorphic males, bombykal never elicited responses from neurones in the AL innervated by the grafted female antenna, although neurones in the control AL responded normally to pheromonal stimulation of the male's own antenna. Injection of cobaltous ions at the completion of recording was accomplished by passing 10 nA depolarizing current pulses of 200–300 ms duration for 15 min at 1 Hz.

group 1). In the brains of male gynandromorphs, the MGC was always missing from the AL innervated by the female graft (Fig. 1c). Conversely, the brains from female gynandromorphs invariably exhibited a structure resembling the MGC in the AL innervated by the male graft (Fig. 1a). In 100 control larvae, we transplanted a disk of the same sex into a host (Table 1, group 2) or re-implanted its own imaginal disk (Table 1, group 3). Of these animals, 86% survived; in 53% of the survivors, the graft formed an antenna; and in 72% of these the operated AL was innervated by the antennal nerve from the graft. Under the light microscope, the ALs of the latter animals were indistinguishable from those of unoperated animals: a male AL innervated by a male graft formed a MGC, but a female AL innervated by a female graft did not. We have observed that the degree of innervation of the AL by grafts ranges from none (where the antennal nerve does not reach the AL) to complete (where there is a full-size antennal nerve). Even in ALs

innervated by reduced numbers of male antennal axons, a readily identifiable MGC-like structure developed.

With the use of intracellular recording and staining⁸, we characterized neurones in the ALs of gynandromorphic moths. In 14 of 16 females with ALs innervated by male antennal grafts, male-like interneurons were present (21 out of 34 cells tested) that are never observed in normal females. Intracellular recordings from such cells (Fig. 2) showed that they responded postsynaptically when the transplanted antenna was stimulated with bombykal. Intracellular staining with cobalt sulphide revealed, moreover, that these cells sent dendritic processes into the MGC-like region of neuropil (Fig. 3). In four gynandromorphic males, whose operated ALs received sensory inputs from transplanted female antennae (lacking pheromone receptors), the characteristic male interneurons were not found. The MGC was missing in these ALs, and no interneurons were observed that either responded when the transplanted antenna was stimulated with bombykal (11 cells tested) or had neurites in the region of the AL where the MGC would normally form. We recorded intracellularly from AL neurones in four males with the same-sex grafts (from group 2 of Table 1) and one male whose own antennal disk had been re-implanted (from group 3 of Table 1). We observed AL neurones that were responsive to bombykal (5 out of 13 cells tested) and had arborizations in the MGC. Finally, normal females and males have AL neurones that respond when the antenna is stimulated with *trans*-2-hexenal ('leaf aldehyde') but not with bombykal, and these cells have arborizations in 'ordinary' glomeruli but not in the MGC⁸. We have observed analogous cells in female and male gynandromorphs. Thus, we note a positive correlation in AL neurones between physiological responsiveness on stimulation of the antenna with pheromone and possession of dendritic arborizations in the MGC.

These observations reveal a remarkable degree of plasticity in the development of sexually dimorphic neurones in the moth's AL. We find that the gender of afferent axons has a pronounced role in the differentiation of AL neurones. A MGC-like structure develops only in ALs, female or male, innervated by axons from a grafted male antenna. In female gynandromorphs, certain AL neurones show masculine patterns of arborization in the MGC-like structure. Moreover, these neurones establish functional connections with axons from the grafted male antenna and respond postsynaptically when the antenna is stimulated by the female pheromone. Whether the axons from the grafted antenna of one sex serve to sustain or suppress the development of gender-specific interneurons in the ALs of their host of the opposite sex, or to remodel or induce response properties and patterns of dendritic branching

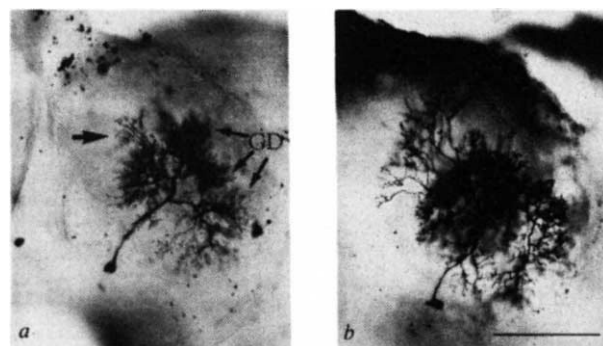


Fig. 3 a, Gynandromorphic female interneurone from which the intracellular recording in Fig. 2 was obtained. The cell was stained intracellularly with cobaltous sulphide, intensified with silver and viewed frontally in a whole-mount⁸. Like the normal male-specific, type-II local interneurone⁸ shown in b, the cell in a lacks an obvious axon and has dendrites (GD) ramifying within the glomerular neuropil and within the MGC-like structure (arrow). The cell body is located at the end of the primary neurite in the lateral cell cluster. Scale bar, 200 μm .

Table 1 Summary of antennal disk-implantation experiments

Group	Type of disk grafted	Total animals*		Everted graft		Innervated AL†	
		Female	Male	Female	Male	Female	Male
1	Opposite sex	163 (122)	166 (135)	65	60	44 (12)	30 (13)
2	Same sex	13 (12)	34 (29)	8	14	6 (4)	10 (5)
3	Self	19 (18)	34 (27)	13	11	10 (7)	7 (4)

* Values in parentheses indicate number of animals that survived surgery.

† Values in parentheses indicate number of ALs sectioned for light-microscopic examination.

characteristic of the graft's gender, is an intriguing but unanswered question. Its answer will lead to a better understanding of the mechanisms underlying development of sexually dimorphic components in the CNS, and possibly those underlying developmental plasticity of neurones in general.

We thank Drs I. D. Harrow and L. P. Tolbert for critical reading of the manuscript; M. Imperato, R. Montague, A. Papier, P. Quartararo and J. Torrisi for assistance; Drs A. H. Baumhover, J. A. Svoboda and C. M. Williams for supplying *Manduca* eggs; and Dr K. H. Dahm for providing bombykal. This work was supported by USPHS research grants AI-16150 and AI-17711, NSF grants BNS-77-13281 and BNS-80-13511, USPHS training grant NS-07112-04 (A.M.S. and S.G.M.) and an NSF predoctoral fellowship (A.M.S.).

Received 27 April; accepted 16 June 1982.

- Boeckh, J., Boeckh, V. & Kuhn, A. *Olfaction Taste*, Paris 6, 315–321 (1977).
- Boeckh, J. & Boeckh, V. *J. comp. Physiol.* **132**, 235–242 (1979).
- Camazine, S. & Hildebrand, J. *Soc. Neurosci. Abstr.* **5**, 155 (1979).
- Hausen, K. & Strausfeld, N. *J. Proc. R. Soc. B208*, 57–71 (1980).
- Hildebrand, J. G. *et al.* in *Insect Neurobiology and Pesticide Action* (ed. Rickett, F. E.) 375–382 (Society for Chemical Industry, London, 1980).
- Strausfeld, N. *J. Nature* **283**, 381–383 (1980).
- Wohlens, D. & Bacon, J. *Cell Tissue Res.* **209**, 371–382 (1980).
- Matsumoto, S. G. & Hildebrand, J. G. *Proc. R. Soc. B213*, 249–277 (1981).
- Prillinger, L. *Cell Tissue Res.* **215**, 563–575 (1981).
- Allen, N. & Hodge, C. R. *J. econ. Ent.* **48**, 526–528 (1955).
- Allen, N., Kinard, W. S. & Jacobson, M. J. *J. econ. Ent.* **55**, 347–351 (1962).
- Sanes, J. R. & Hildebrand, J. G. *Dev. Biol.* **51**, 300–319 (1976).
- Sanes, J. R. & Hildebrand, J. G. *Dev. Biol.* **51**, 282–299 (1976).
- Jawłowski, H. *Annls Univ. Mariae Curie, Skłodowska C8*, 403–434 (1954).
- Tolbert, L. P. & Hildebrand, J. G. *Proc. R. Soc. B213*, 279–301 (1981).
- Sanes, J. R. & Hildebrand, J. G. *Wilhelm Roux's Arch. dev. Biol.* **178**, 71–78 (1975).
- Sanes, J. R., Prescott, D. J. & Hildebrand, J. G. *Brain Res.* **119**, 389–402 (1977).
- Hildebrand, J. G., Hall, L. M. & Osmond, B. C. *Proc. natn. Acad. Sci. U.S.A.* **76**, 499–503 (1979).
- Schneiderman, H. A. in *Methods in Developmental Biology* (eds Wilt, F. H. & Wessels, N. K.) 753–766 (Crowell, New York, 1967).

Exaggerated salt appetite of spontaneously hypertensive rats is decreased by central angiotensin-converting enzyme blockade

R. DiNicolantonio, J. S. Hutchinson & F. A. O. Mendelsohn

University of Melbourne, Department of Medicine, Austin Hospital, Heidelberg, Victoria 3084, Australia

Injection of angiotensin II (AII) into the cerebral ventricles at doses as low as 1 pmol h⁻¹ results in a marked stimulation of salt and water ingestion in the rat^{1,2}. Evidence that AII is produced in the central nervous system independently of the circulating renin-angiotensin system (RAS)^{3,4} raises the possibility that endogenous brain AII is involved in the physiological regulation of thirst⁵. The role of brain AII in salt appetite is still unclear⁶. Here we confirm that the spontaneously hypertensive rat (SHR), believed to have elevated levels of brain AII⁷, possesses an exaggerated salt appetite compared with its normotensive controls. We also show that this exaggerated salt appetite is reduced by chronic central treatment with the angiotensin-converting enzyme inhibitor, captopril, while that of the normotensive controls is unaffected. Our study suggests that a central neuropeptide, probably AII, is involved in the maintenance of the exaggerated salt appetite in this model of hypertension.

Intracerebroventricular (i.c.v.) administration of AII receptor antagonists⁸ or angiotensin-converting enzyme inhibitors⁹ to the spontaneously hypertensive rat of the Okamoto strain, but not to the normotensive control Wistar Kyoto (WK) rat results in a marked, rapid, hypotensive response independent of the circulating RAS. Furthermore, an increased density of

immunoreactive AII has been reported in hypothalamic neurones of the SHR¹⁰. As the circulating RAS in the SHR does not differ from that in the WK rat¹¹, these findings suggest an overactivity of the brain RAS in the SHR. The SHR have an exaggerated salt appetite compared with WK rats¹². This, given the evidence for overactivity of the brain RAS in the SHR, may be due to overproduction of central AII. To test this hypothesis, we have monitored fluid intake in groups of SHR and WK animals offered a choice of water or saline before and after chronic infusion of captopril into the left lateral cerebral ventricle in an attempt to block angiotensin II production.

Figures 1 and 2 show the fluid intakes of SHR and WK rats offered a choice of saline or water before and after infusion of either captopril, 80 µg h⁻¹, or vehicle. In the pretreatment control period, SHR drank 23.3 ± 1.0 ml per 100 g body weight per day (mean ± s.e.m. of 100 observations) of saline and 4.8 ± 0.3 ml per 100 g per day (*n* = 100) of water; this represents a saline preference of 83% of total fluid intake. In the same period, WK rats drank 7.1 ± 0.5 ml per 100 g per day (*n* = 100) of saline and 9.3 ± 0.3 ml per 100 g per day (*n* = 100) of water, saline representing 43% of fluid intake. Therefore, SHR had nearly double the total fluid intake of the WK rats (*t* = 8.3, *P* < 10⁻¹³) and also a significantly higher proportion of saline in this intake (*t* = 12.7, *P* < 10⁻²⁴). The effect of central captopril on the intake and choice of fluid of SHR is shown in Fig. 1. Vehicle-treated SHR showed no change in saline intake during the infusion period compared with the pre-implantation period (*F*_{1,120} = 0.3, *P* > 0.6). In contrast, captopril-infused rats showed a significant decrease (23%) in saline intake during the treatment period compared with either the pre-implantation period (*F*_{1,108} = 20.3, *P* < 10⁻⁴) or vehicle-infused rats (*F*_{1,112} = 9.2, *P* < 3 × 10⁻³). Captopril-treated SHR also showed an increased (172%) water intake compared with the vehicle-infused SHR during the treatment period (*F*_{1,112} = 33.9, *P* < 10⁻⁷), while no significant differences existed between the two groups during the pretreatment period (*F*_{1,112} = 2.5, *P* > 0.12). Whereas saline represented 83 ± 1% of total fluid intake of SHR before treatment, this fell to 77 ± 2% after captopril (*F*_{1,112} = 6.4, *P* < 0.01). Vehicle-treated SHR showed an increase of proportional saline intake from 84 ± 1% to 88 ± 1% (*F*_{1,126} = 13.7, *P* < 3 × 10⁻⁴). In contrast, as shown in Fig. 2, WK rats showed no changes in saline intake, water intake or proportional saline intake with either captopril or vehicle treatment.

The food intake of captopril-infused SHR did not differ from that of vehicle-infused SHR either before (68 versus 67 g per day, respectively; *F*_{1,80} = 0.1, *P* > 0.8) or during i.c.v. treatment (70 versus 68 g per day; *F*_{1,80} = 0.9, *P* > 0.3). Similarly, the food intake of captopril-infused WK rats was not different from that of vehicle-infused WK rats either before (66 versus 64 g per day, respectively; *F*_{1,56} = 1.6, *P* > 0.2) or during i.c.v. treatment (64 versus 61 g per day; *F*_{1,70} = 1.4, *P* > 0.2). The growth rate of captopril-infused SHR did not differ from that of vehicle-infused SHR either before (0.4 versus 0.01 g per day, respectively; *F*_{1,80} = 0.1, *P* > 0.8) or during i.c.v. treatment (2.0 versus 2.1 g per day; *F*_{1,80} = 0.05, *P* > 0.8). Similarly, the growth rates of captopril- or vehicle-infused WK rats were also similar before and during treatment (0.01 versus 0.5 g per day, respectively; *F*_{1,56} = 0.6, *P* > 0.4; and 0.4 versus 0.8 g per day, respectively; *F*_{1,70} = 0.1, *P* > 0.8). There was no significant difference in total fluid intake between either central captopril- or vehicle-infused SHR during the treatment period (24 versus 26 ml per 100 g per day, respectively; *F*_{1,112} = 2.3, *P* > 0.1).

It is possible that the fall in salt appetite during central captopril was secondary to renal sodium retention due to a fall in blood pressure. However, the animals showed no change in blood pressure during chronic infusion of captopril at this dose (pre-infusion: 187 ± 5 mmHg, post-infusion: 188 ± 4 mmHg; *F*_{1,60} = 0.02, *P* > 0.5). To examine this possibility further, five male SHR were treated with hydralazine (5 mg per kg per day, orally) over a 4-day period. Blood pressure, measured by a tail cuff method, fell from a control value of 210 ± 10 mmHg to

149 ± 4 mmHg on day 4 of treatment. Saline consumption was significantly depressed on day 1 of treatment compared with the pretreatment period (14.1 versus 24.7 ml per 100 g per day, respectively; $F_{1,20} = 8.8$, $P < 0.01$), but this was not maintained and there was no difference in saline consumption on the following 3 days of treatment relative to the pretreatment period (22.3 versus 24.7 ml per 100 g per day; $F_{1,28} = 2.2$, $P > 0.1$), despite a maximal fall in blood pressure at this time.

To examine the potency of the effect of captopril on the salt appetite of the SHR, the experiments were repeated using a 10-fold lower dose of captopril ($8 \mu\text{g h}^{-1}$) infused into the lateral cerebral ventricle over a 7-day period in nine adult male SHR.

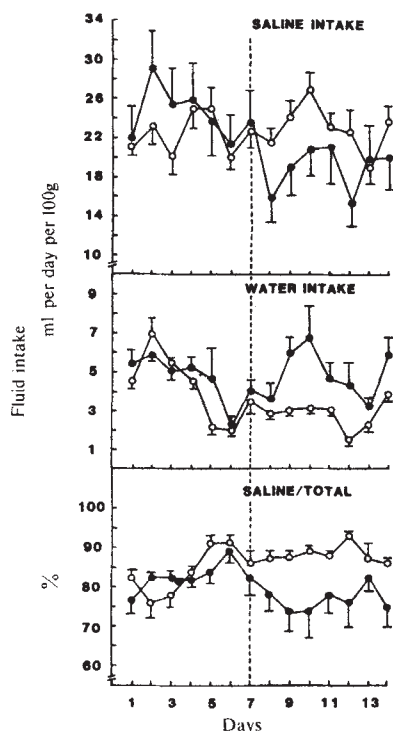


Fig. 1 Fluid intake of male SHR offered a choice of 0.9% saline or tap water. We used 20 male SHR and 20 male WK rats (weight range 320–410 g) housed in individual metabolic cages in conditions of 12-h light and 12-h dark (lights on 0700–1900 h), with room temperature maintained at 22 °C. Standard pelleted laboratory chow (Clarke King GR 2+, Melbourne) was available *ad libitum* and daily food intake was estimated by weighing uneaten and spilt food. A choice of tap water and 0.9% saline was available from individual bottles fitted with dripless stainless steel spouts (Gelman Clemco, Melbourne). Fluid intake was measured daily at 1700 h by weighing the bottles and the positions of saline and water bottles in individual cages were alternated daily. Following a 4-day acclimatization period, a further 7-day period was used to obtain control values of fluid and food intakes. Daily body weight was recorded at 1700 h and fresh saline prepared every second day. Following this control period, animals were lightly anaesthetized with ether and a 27-gauge stainless steel cannula was implanted into the left lateral cerebral ventricle at coordinates AP 1.0, L 1.5, H 4.0 using the bregma as reference. Cannulae were attached by PE 60 tubing to an Alzet osmotic minipump (Alza Corporation), with 10 SHR and 10 WK rats receiving captopril at $80 \mu\text{g h}^{-1}$ and 10 SHR and 10 WK animals receiving artificial cerebrospinal fluid as the vehicle. The minipump was inserted subcutaneously into the interscapular region. Following surgery, the animals were returned to their individual cages and the measurements continued for a further 7 days. At the end of this period, rats were decapitated and post-mortem examination of the brain performed to verify correct placement of the cannulae. The significance of changes in fluid intake was assessed by two-way analysis of variance. Fluid intakes were expressed per 100 g of body weight and the saline preference represents the percentage saline intake as a proportion of the total fluid intake. The dotted line indicates the start of the i.c.v. infusion of either captopril (●) or vehicle (○).

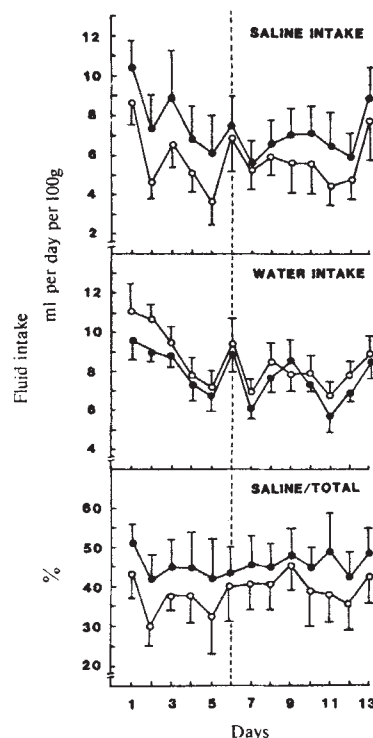


Fig. 2 Fluid intake of male WK rats offered a choice of 0.9% saline or tap water (see Fig. 1 legend). ●, Rats given i.c.v. captopril; ○, vehicle-treated rats.

This dose of captopril produced no change in blood pressure over the treatment period when compared with the 7-day control period (175 ± 4 versus 178 ± 4 mmHg, respectively; $F_{1,48} = 0.3$, $P > 0.5$). Mean saline consumption was 37.2 ± 2.2 ml per 100 g per day during the control period compared with 22.4 ± 1.0 ml per 100 g per day during the last 4 days of captopril infusion ($F_{1,88} = 22.3$, $P < 10^{-5}$). No change in saline intake occurred in nine vehicle-infused SHR during this period when compared with the pretreatment period (28.1 ± 2.0 versus 31.2 ± 2.0 ml per 100 g per day, respectively; $F_{1,88} = 1.1$, $P > 0.7$).

The selective attenuation of the exaggerated salt appetite of the SHR by central captopril, as shown here, suggests that elevated central AII may contribute to this elevated salt appetite. The fact that central captopril had no effect on the food intake or growth rate of the SHR, and that central captopril was without any effect in the WK rat argues against a nonspecific effect of the drug. Furthermore, whereas a specific decrease in the proportional saline intake was seen in SHR treated centrally with captopril, total fluid intake remained unchanged, suggesting a specific action on sodium appetite.

The possibility that blockade of the circulating RAS, following leakage of captopril from central sites, contributed to the effects seen here may be excluded, as peripherally administered captopril not only lowers the threshold for salt detection but also increases saline ingestion in the rat¹³.

As bradykinin^{14,15} and the enkephalins¹⁶ are substrates for angiotensin-converting enzyme, it is possible that the effects of captopril in our study are mediated by changes in these or other neuropeptides. However, the opioids seem to stimulate ingestive behaviour^{17,18} and elevation of their central levels by captopril would not explain the results observed. According to one study, central administration of bradykinin failed to cause drinking¹⁹, although a large dose of bradykinin has been reported to attenuate AII-induced drinking²⁰; potentiation of this peptide might therefore contribute to the actions of central captopril in the SHR.

Centrally administered saralasin did not decrease thirst²¹ or salt appetite⁶ in previous reports. However, saralasin has agonist activity in most systems²² and actually potentiated drinking in

one study²³. Furthermore, this peptide is of relatively high molecular weight and it has been suggested that captopril may have access to sites inaccessible to saralasin²⁴.

The cerebrospinal fluid of hypertensive humans has been reported to contain elevated levels of angiotensin I²⁵ and angiotensinogen²⁶. It is intriguing to note, therefore, that hypertensives show a preference for sodium chloride and exhibit increased daily fluid intake²⁷, a pattern strikingly similar to that observed in the SHR. The present findings therefore raise the possibility that the brain RAS may be hyperactive in some forms of human essential hypertension.

This work was supported by the National Health and MRC and the National Heart Foundation of Australia.

Received 12 March; accepted 17 June 1982.

1. Bryant, R. W., Epstein, A. N., Fitzsimons, J. T. & Fluharty, S. J. *J. Physiol., Lond.* **301**, 365–382 (1980).
2. Avrih, D. B. & Fitzsimons, J. T. *J. Physiol., Lond.* **301**, 349–364 (1980).
3. Phillips, M. I. *Neuroendocrinology* **25**, 354–377 (1978).
4. Okamura, T., Clemens, D. L. & Inagami, T. *Proc. natn. Acad. Sci. U.S.A.* **78**, 6940–6943 (1981).
5. Ganten, D., Fuxe, K., Phillips, M. I., Mann, J. F. E. & Ganten, U. *Front. Neuroendocr.* **5**, 61–99 (1978).
6. Coghlan, J. P. *et al. Science* **214**, 195–197 (1981).
7. Phillips, M. I. in *Nervous System and Hypertension* (eds Meyer, P. & Schmitt, H.) 102–105 (Wiley, New York, 1979).
8. Phillips, M. I. *et al. Nature* **270**, 445–447 (1977).
9. Hutchinson, J. S., Mendelsohn, F. A. O. & Doyle, A. E. *Hypertension* **2**, 546–550 (1980).
10. Phillips, M. I., Weyhenmeyer, J. & Ganten, D. in *Proc. 4th int. Symp. Rats with Spontaneous Hypertension and Related Studies*, Heidelberg (Abstr.) 21 (1981).
11. Phillips, M. I. in *Enzymatic Release of Vasoactive Peptides* (eds Gross, F. & Vogel, G.) 335–363 (Raven, New York, 1980).
12. McConnell, S. D. & Henkin, R. I. *Am. J. Physiol.* **225**, 624–627 (1973).
13. Fregly, M. J. *Pharmac. exp. Ther.* **215**, 407–412 (1980).
14. Erdos, E. G. *Circulation Res.* **36**, 247–255 (1975).
15. Dorer, F. E., Kahn, J. R., Lentz, K. E., Levine, M. & Skeggs, L. T. *Circulation Res.* **34**, 824–827 (1974).
16. Erdos, E. G., Johnson, A. R. & Boyden, N. T. *Biochem. Pharmac.* **27**, 843–848 (1978).
17. McKay, L. D., Kenney, N. J., Edens, N. K., Williams, R. H. & Woods, S. C. *Life Sci.* **29**, 1429–1434 (1981).
18. Brown, D. R. & Holtzman, S. G. *Life Sci.* **28**, 1287–1294 (1981).
19. Fitzsimons, J. T. *J. Physiol., Lond.* **214**, 295–303 (1971).
20. Kenney, N. J. in *Neural Mechanisms of Goal Directed Behaviour and Learning* (eds Thompson, R. F., Hicks, L. H. & Shvyrkov, V. B.) 437–446 (Academic, New York, 1980).
21. Lee, M.-C., Thrasher, T. N. & Ramsay, D. J. *Am. J. Physiol.* **240**, R75–R80 (1981).
22. Marshall, G. R. *Fedn Proc.* **35**, 2494–2501 (1976).
23. Malvin, R. L., Mow, D. & Vander, A. J. *Science* **197**, 171–173 (1977).
24. Antonaccio, M. J. & Kerwin, L. *Hypertension* **3**, Suppl I, 154–162 (1981).
25. Finkelman, S., Fischer-Ferraro, C., Diaz, A., Goldstein, D. J. & Nahmod, V. E. *Proc. natn. Acad. Sci. U.S.A.* **69**, 3341–3344 (1972).
26. Ito, T. *et al. Hypertension* **2**, 432–436 (1980).
27. Schechter, P. J., Horwitz, D. & Henkin, R. I. *J. Am. med. Ass.* **225**, 1311–1315 (1973).

Interphotoreceptor retinol-binding proteins: possible transport vehicles between compartments of the retina

Y. L. Lai*, B. Wiggert†, Y. P. Liu† & G. J. Chader†

* Bethesda Eye Institute, St Louis, Missouri 63110, USA

† National Eye Institute, National Institutes of Health, Bethesda, Maryland 20205, USA

In the eye, vitamin A (retinol) is mainly stored in the retinal pigment epithelium (RPE) although its primary function is in the visual process in the photoreceptor organelles of the neural retina (NR). It is well established that during light adaptation, the amount of retinol drops in the NR but rises in the RPE¹. During dark adaptation, the converse occurs. This indicates a migration of retinoid between the two tissues, the direction of which is dictated by the state of light or dark adaptation of the photoreceptors. The mechanism by which this migration is effected is unknown. We now present evidence that at least one protein exists in the subretinal space or on the cell surfaces which demonstrates many of the unique characteristics one would expect of an interphotoreceptor retinol-binding protein and could function as a vitamin A transport vehicle between NR and RPE.

A technique in which sub-retinal fluid is collected by washing the inner photoreceptor space of the detached retina and by

irrigating the apical surface of the RPE and the sub-choroidal space has made possible the study of biochemical parameters involved in the dynamic interaction of these tissues². Previously, gel electrophoresis and scanning electron microscopy have given strong evidence that the tissues do, in fact, remain intact during the lavage procedure². We have now measured ³H-retinol binding in four retinal compartments in the rabbit eye: (1) supernatant of homogenized neural retina, (2) washes of the photoreceptor surface (retinal fraction of the sub-retinal fluid space (SRF/NR)), (3) washes of the apical surface of the retinal pigment epithelium (RPE fraction of the sub-retinal space (SRF/RPE)) and (4) supernatant of homogenized RPE-choroidal tissue. Eyes from young adult New Zealand albino rabbits were removed under light anaesthesia, NR and RPE gently separated and the surfaces washed with isotonic saline as previously described². Animals were light- or dark-adapted overnight before experimentation. Manipulations involving dark-adapted samples were performed under dim red light. Preparation of the supernatant fraction of tissue homogenates and gradient analysis techniques have already been described³.

We have previously shown that a 7S retinol-binding protein is present in the supernatant fraction of adult retina of several species as well as the more ubiquitous 2S cellular retinol-binding protein, CRBP^{4,5}. Retinol is bound endogenously to the 7S species as well as to the 2S protein, as assessed by relative fluorescence intensity measurements⁷. The 7S species appears to be primarily associated with the photoreceptor outer segments since (1) it is not apparent before outer segments appear in the embryonic retina⁶, (2) it is not present when outer segments are absent, as in the disease retinitis pigmentosa⁷ and (3) its concentration is increased when outer segments are partially purified although it is lost on extensive purification⁶. Moreover, the 7S retinol-binding activity is greater in the soluble extract of light-adapted frog retina and outer segments than in those of the dark-adapted state⁸. In the present study with rabbit, we observed both 7S and 2S ³H-retinol-binding proteins in the retinal supernatant fraction after sucrose gradient analysis (results not shown). Both binding proteins (7S at about fraction 13; 2S at about fraction 28) can also be observed in the washings of the photoreceptor surface of the retina, the subretinal fluid area (SRF/NR; Fig. 1a, b). 7S binding is appreciably greater (three times) in the light-adapted state than in the dark-adapted while the amount of 2S binding is virtually identical in the two samples (Table 1).

The specific activities of the 7S and 2S binding are much higher in the subretinal compartments than in the supernatant fraction of retinal homogenate (Table 1). In the light, this amounts to a 24-fold difference in 7S and a 65-fold difference in 2S binding when the SRF/NR wash is compared with the retinal supernatant. Examination of the washings of the pigment epithelial surface (SRF/RPE) reveals a similarly dramatic light/dark difference in 7S receptor binding, although there is a pronounced light effect on 2S binding as well which is not observed in the SRF/NR wash (Fig. 1c, d; Table 1). 7S binding is 6.6-fold higher in the light than the dark while 2S binding is 4.7-fold higher in the light. 2S ³H-retinol binding is present in the supernatant fraction of RPE-choroidal tissue homogenate

Table 1 ³H-retinol binding in retinal compartments

Compartment	pmol ³ H-retinol bound per mg protein			
	7S		2S	
	Light	Dark	Light	Dark
Retinal supernatant	3.4 ± 0.5	2.1 ± 0.4	1.5 ± 0.3	2.6 ± 0.5
SRF/NR wash	83.0 ± 10	28.0 ± 5	97.0 ± 12	90.0 ± 10
SRF/RPE wash	19.2 ± 4	2.9 ± 0.6	198.0 ± 25	42.6 ± 7
RPE choroidal supernatant	*	*	4.0 ± 0.6	4.6 ± 0.8

Conditions as given in Fig. 1 legend. Values given are averages ± s.d. of duplicate gradients from three separate experiments.

* Below the limits of detection.

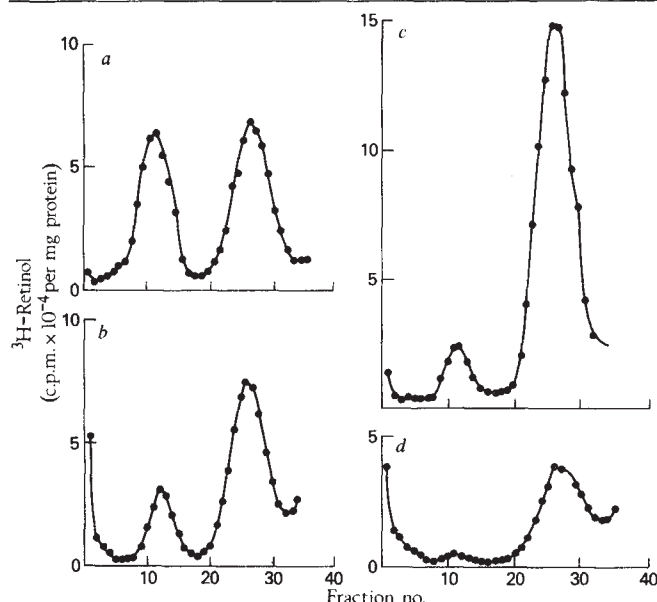


Fig. 1 Sucrose density gradient centrifugation patterns of ^3H -retinol binding to washings of *a, b*, the retinal photoreceptor surface (SRF/NR); *c, d*, the pigment epithelial surface (SRF/RPE). *a, c*, From light-adapted animals; *b, d*, from dark-adapted animals. The saline washes of the tissue surfaces were diluted 1:2 with Tris buffer (10 mM Tris, 10 mM KCl, 1 mM EDTA, 1 mM dithiothreitol) before gradient analysis and then incubated with 2.8×10^{-7} M all-*trans*, 1- ^3H -retinol (5 Ci mmol $^{-1}$; NEN) for 2 h at 4°C in the dark. 250 μl aliquots were then layered on 4.6 ml continuous sucrose gradients (5–20% sucrose) in the Tris buffer and centrifuged for 16 h at 243,000 g. About 35 fractions were collected and their radioactivity determined. The top of the gradient is at higher fraction numbers. Protein concentration was determined by the method of Lowry *et al.*¹⁵. Binding specificity was checked by incubating parallel samples with a 500-fold molar excess of unlabelled retinol. The total amount of specific binding in each peak was determined by summing the total c.p.m. under the peak area and then subtracting the non-specific counts which lay below the peak nadirs. Patterns given are representative patterns from three separate experiments, each having duplicate analysis within the separate experiment.

but has a much lower specific activity than in either the SRF/NR or SRF/RPE fractions (Table 1). There is no detectable 7S binding in this compartment.

Kuhn^{9,10} has shown that several proteins are differentially bound to photoreceptor membranes in light and dark. This offers the simplest explanation for our results—that the 7S and perhaps the 2S species are 'peripheral' proteins, loosely bound at the membrane surface. The enhancement in binding in the SRF washes compared with the tissue homogenates and the light/dark effect are both consistent with this postulate. As ^3H -retinoid binding studies do not measure the actual amount of binding protein, however, but rather the number of unoccupied binding sites, our results could be affected by several factors. For example, the decreased ^3H -retinol binding observed at 7S in the SRF/NR and SRF/RPE compartments and at 2S in the SRF/RPE compartment following dark adaptation could be due to an increase in endogenous unlabelled retinol in these compartments, a decrease in the actual amount of binding protein (turnover) or a change in the binding properties of the proteins. It will be of interest to investigate this in subsequent studies. Whether either of these proteins exhibits enzymatic activity is unclear but should also be investigated in light of the interesting finding that a defect in vitamin A metabolism may be involved in hereditary retinal degeneration in an animal model of retinitis pigmentosa¹¹. LaVail *et al.*¹² have also established striking differences in inter-photoreceptor matrix mucosubstances between normal rats and those with inherited retinal dystrophy. What is clear from the present study is that the specific activities of the 7S protein and the 2S CRBP are

extremely high in the extracellular subretinal space and that the apparent binding activities are markedly affected by light.

Since free retinol is labile and a potentially toxic substance which can alter cellular membranes¹³, as exemplified by retinoid-evoked release of enzymes from NR and RPE lysosomes¹⁴, some sort of carrier might be expected to be present at the various stages of retinoid transit into and through the cells. Moreover, retinoid, used in the visual cycle of the retina, is supplied through the PE cell from the choroidal blood supply, necessitating a generalized vectorial movement of retinoid as well as the more localized, light-evoked 'shuttle' movement between PE and retina during the visual cycle. It would seem likely, then, that the 7S binding protein and the 2S CRBP are both likely candidates for the role of retinoid transport vehicle between the sensory retina and the pigment epithelium. The 7S protein which we call interphotoreceptor retinol-binding protein seems to be a particularly interesting protein for further study due to its restricted presence in the retina, its localization and high concentration in the subretinal photoreceptor matrix and at the membrane surface and its apparent control of binding by light.

Received 23 March; accepted 24 June 1982.

1. Dowling, J. *Nature* **168**, 114–118 (1960).
2. Lai, Y.-L. *et al. Expl Eye Res.* **34**, 7–14 (1982).
3. Wiggert, B. & Chader, G. *Expl Eye Res.* **21**, 143–151 (1975).
4. Wiggert, B., Bergsma, D. & Chader, G. *Expl Eye Res.* **22**, 411–418 (1976).
5. Wiggert, B., Bergsma, D., Lewis, M. & Chader, G. *J. Neurochem.* **29**, 947–954 (1977).
6. Wiggert, B., Mizukawa, A., Kuwabara, T. & Chader, G. *J. Neurochem.* **30**, 653–659 (1978).
7. Bergsma, D., Wiggert, B., Funahashi, M., Kuwabara, T. & Chader, G. *Nature* **265**, 66–67 (1977).
8. Wiggert, B., Derr, J., Fitzpatrick, M. & Chader, G. *Biochim. biophys. Acta* **582**, 115–121 (1979).
9. Kuhn, H. *Biochemistry* **17**, 4389–4395 (1978).
10. Kuhn, H. *Nature* **238**, 587–589 (1980).
11. Berman, E., Segal, N., Photiou, S., Rathman, H. & Feeney-Burns, L. *Nature* **293**, 217–220 (1981).
12. LaVail, M., Pinto, L. & Yasumura, D. *Invest. Ophthalmol. Vis. Sci.* **21**, 658–668 (1981).
13. Meeks, R., Zaharevitz, D. & Chen, R. *Archs Biochem. Biophys.* **207**, 141–147 (1981).
14. Dewar, A., Barron, G. & Reading, H. *Expl Eye Res.* **20**, 63–72 (1975).
15. Lowry, O., Rosebrough, N., Farr, A. & Randall, R. *J. biol. Chem.* **193**, 265–275 (1953).

Detection of K^+ and Cl^- channels from calf cardiac sarcolemma in planar lipid bilayer membranes

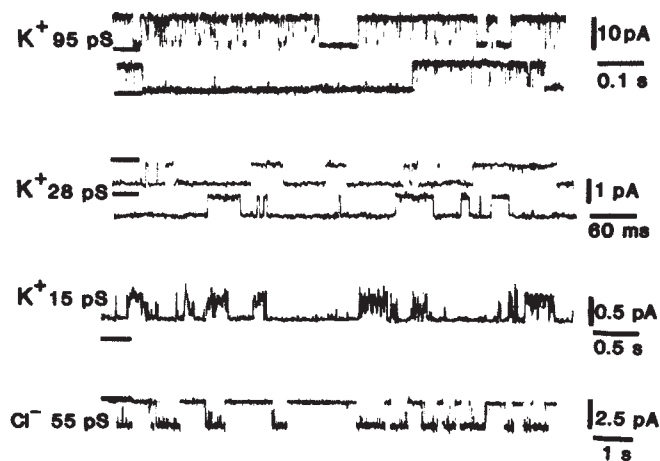
Roberto Coronado & Ramon Latorre

Department of Physiology and Biophysics, Harvard Medical School, Boston, Massachusetts 02115, USA

The ionic currents underlying the cardiac action potential are believed to be much more complex than those in nerve^{1,2}. During the cardiac action potential, various membrane channels control the flow of K^+ , Na^+ , Ca^{2+} and Cl^- across the sarcolemma of cardiac muscle cells^{2–7}. Thus, it has become increasingly clear that a detailed knowledge of the mechanisms that activate (or inactivate) heart channels is required to understand cardiac excitability^{8,9}. We report here the use of planar lipid bilayer techniques to detect and characterize K^+ and Cl^- channels in purified heart sarcolemma membrane vesicles. We have identified four different types of channel on the basis of their selectivity, conductance and gating kinetics. We present in some detail the properties of a K^+ channel and a Cl^- channel. We have tentatively identified the K^+ channel with the i_t type of current found in Purkinje¹⁰, myocardial ventricular¹¹ and atrial¹² fibres. The chloride channel might be related to the transient chloride current found in Purkinje fibres^{5,6,13}.

We have used sarcolemmal membranes isolated from calf myocardium by a method similar to that described in ref. 14. This method is particularly suitable for our purposes because it minimizes contamination by sarcoplasmic reticulum. Ionic channels present in membrane vesicles purified from heart

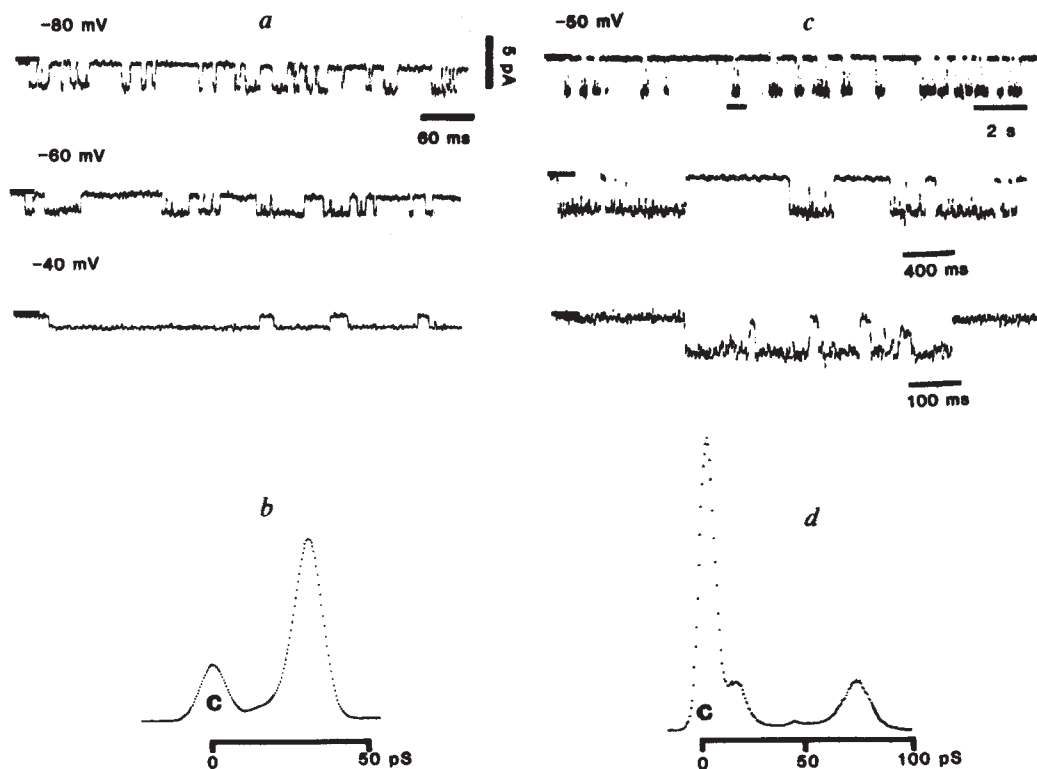
Fig. 1 Calf cardiac channels incorporated into planar lipid bilayers. Planar lipid bilayer membranes were formed from a mixture of bovine brain phosphatidylethanolamine (PE) and bovine brain phosphatidylserine (PS) at a PE/PS ratio of 1:1 in decane. This solution was made at a concentration of 50 mg ml⁻¹. Lipid bilayers were formed in Teflon partitions with apertures of 5×10^{-4} cm². The bilayer separated two aqueous chambers, each with a volume of ~ 100 μ l. Electrical measurements were made with a two-electrode voltage clamp¹⁷. One of the chambers (*cis* side) was connected to a function generator to control membrane potential, while the second chamber (*trans* side) was connected to a current-to-voltage converter. Thus, the *trans* side is virtual ground. A positive current is defined as cation flow into the *trans* side. Ag/AgCl electrodes in series with 0.2 M KCl agar bridges served to connect the bilayer solutions to the electronics. Cardiac sarcolemma was prepared from calf ventricle using the procedure of Jones *et al.*¹⁴. Membranes were stored at -40°C in 0.4 M sucrose, 10 mM HEPES-Tris, pH 7.4. The protocol used to incorporate single channels into the bilayer was as follows. Thawed vesicles were diluted to a final concentration of 50–100 μ g protein per ml in 0.1 M KCl, 1 mM CaCl₂, 5 mM HEPES-Tris, pH 7.2. After forming the bilayer in symmetrical 0.1 M KCl, 5 mM HEPES-Tris, pH 7.2, the *cis* chamber was perfused with the vesicle-containing solution. Once the initial jump in conductance was observed, the *cis* chamber was extensively perfused with 0.1 M KCl, 5 mM HEPES-Tris, pH 7.2, to remove non-incorporated protein. All single-channel records were stored on FM tape for further analysis. By convention, in all records shown, the channel opening is an upward deflection (that is, 95 pS and 15 pS K⁺ channels) if the applied voltage is positive, or a downward deflection if the applied potential is negative (that is, 28 pS K⁺ channel and 55 pS Cl⁻ channel). The conductance of the bare bilayer (≤ 5 pS in all cases) is indicated by the horizontal mark at the beginning of each single-channel trace. For the recordings shown, the applied potential was +100 mV for the 95 pS channel, -40 mV for the 28 pS channel, +30 mV for the 15 pS channel and -50 mV for the 55 pS channel. (K⁺) or (Cl⁻) indicates whether a given type of channel is Nernst selective for K⁺ over Cl⁻ (labelled K⁺) or vice versa (labelled Cl⁻). Experiments were performed at room temperature, $22 \pm 2^\circ\text{C}$.



sarcolemma can be transferred to planar membranes by a Ca²⁺-induced fusion procedure^{15,16}. Incorporation of channels proceeds very quickly, eliciting in a few minutes an increase in bilayer conductance of two or three orders of magnitude provided that two conditions are met: the presence of an osmotic gradient across the bilayer (the vesicle-containing side hyperosmotic), and the presence of Ca²⁺ in the vesicle-containing side. As described elsewhere, these conditions favour fusion of purified membrane vesicles^{15–17} and liposomes¹⁸ to planar bilayers. However, when the solutions bathing the membrane are kept symmetrical, and the Ca²⁺ concentration in the vesicle-containing compartment is low (1 mM), we usually observe the incorporation of single channels of one type. Figure 1 shows

four types of single-channel current (which are also the most common) in the planar membrane. The first three records correspond to channels having reversal potential values typically observed when Cl⁻ is completely excluded. The fourth record corresponds to a channel permeable to Cl⁻ but not K⁺. The 95 pS channel shown in Fig. 1 has a complicated gating kinetics consisting of one open and at least two closed states, is voltage independent and does not discriminate between K⁺ and Na⁺, or among other alkali cations. The 15 pS channel shows a 'noisy' open state and a permeability ratio, P_{K^+}/P_{Na^+} , of 3. None of the channels shown above is activated by Ca²⁺, as they persist even if the Ca²⁺ concentration is 10^{-8} M. The kinetics of the 95 pS channel is, however, substantially modified

Fig. 2 Electrical characteristics of a K⁺ and a Cl⁻ channel. *a*, Recordings of the 28 pS K⁺-conducting channel from the same membrane, at the voltages indicated. 1 KHz filtering. *b*, Histogram showing the relative occurrence of the two conductance levels of the 28 pS channel at -40 mV. It was obtained by feeding long records (>10 s) of current fluctuations, like those in *a*, into a pulse height analyser and sampling the current at a rate of 10^4 s⁻¹. The histogram is built by placing the sampled current in the bin corresponding to that particular current value. Thus, the peaks represent the most frequent values of conductance for a given record. The mean conductance value for the open state is 28 pS. *c*, Recordings of the 55 pS Cl⁻-conducting channel at -50 mV. The bar underlying the single burst in the top trace is expanded in the bottom trace. The bar underlying the last three bursts is expanded in the middle trace. 1 KHz filtering. *d*, Histogram showing the relative occurrence of all current levels present in *c*. All measurements were done on symmetrical 0.1 M KCl, pH 7.2. C indicates the closed state peak.



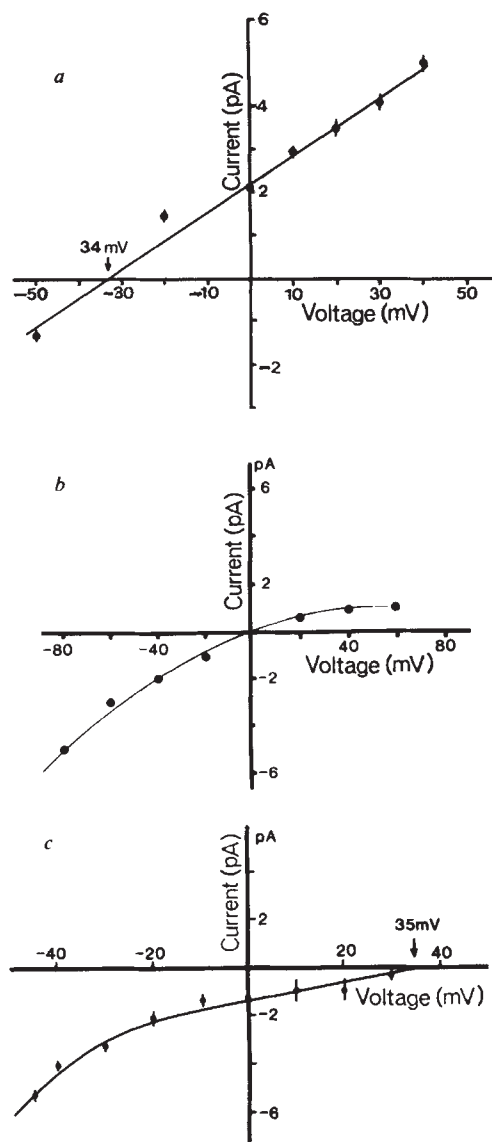


Fig. 3 Single-channel current-voltage curves for the 28 pS K^+ channel and 55 pS Cl^- channel. *a*, Mean single-channel currents obtained from histograms, such as that shown in Fig. 2*b* for the K^+ channel, were plotted for a membrane formed in 0.5 M KCl *cis* side, 0.1 M KCl *trans* side, 5 mM HEPES-Tris, pH 7.2, both sides. *b*, Single-channel current-voltage curve for the Cl^- channel in symmetrical 0.1 M KCl, 5 mM HEPES-Tris, pH 7.2. Time-average currents through the open states were measured by filtering bursts at 5–10 Hz and computing the mean current. *c*, Current-voltage curve for the same channel in 0.5 M KCl *cis* and 0.1 M KCl *trans*. Voltage indicated at the arrow corresponds to the extrapolated zero-current voltage.

by *cis* Ca^{2+} . As the *cis* Ca^{2+} is decreased below 1 mM, the long closing events (clearly seen at the bottom record of the 95 pS channel in Fig. 1) generally disappear. The properties of the 28 pS K^+ channel and the 55 pS Cl^- channel are described in more detail below.

The 28 pS K^+ channel shows well defined rectangular single-channel currents (Fig. 2*a*). The current step size is directly proportional to voltage; however, the time the channel remains in a given conductance state is voltage dependent. As the voltage is made more positive, the channel remains longer in the open configuration. For the records shown in Fig. 2*a*, the mean open time increases from 6 ms at -80 mV to 60 ms at -40 mV. The probability histogram (Fig. 2*b*) further corroborates the two-conductance-state nature of the channel, as only two peaks can be observed. The open state is, however, inter-

rupted by fast closing events which are noticeable at high negative potentials. The mean closed time of these interruptions is 2–4 ms and may indicate the presence of a second kinetically distinguishable closed state. The electrical activity of the Cl^- channel is much more complicated (Fig. 2*c*). The channel may stay closed for several seconds, then it opens and begins to fluctuate rapidly between several conductance states (see bottom record of Fig. 2*c*). We interpret this type of record to represent the fluctuations of a single channel between several open conductance states and at least two closed states. The probability histogram in Fig. 1*d* shows clearly at least three open conductance states with conductances of 15, 39 and 61 pS.

For the K^+ channel shown in Fig. 2*a*, we also measured the current passing through a single channel in the presence of a fivefold KCl concentration gradient (Fig. 3*a*). The reversal potential for the current transported through a single channel is -34 mV, in agreement with that predicted by the Nernst equation for a K^+ -selective channel. We have also measured the bi-ionic potential for 0.1 M KCl (vesicle-containing side)/0.1 M NaCl and found a reversal potential of -40 mV. This implies a permeability ratio $P_{K^+}/P_{Na^+} = 5$.

For Cl^- channels, because of the complexity of the records, we have plotted the average of the weighted mean current through the channel as a function of voltage. Figure 3*b* shows that the mean current through the single channel is a nonlinear function of potential. For negative potentials, current increases more than linearly. For positive potentials, however, the current reaches a limiting value. Analysis of histograms, such as that shown in Fig. 2*d*, indicates that the shape of the curve in Fig. 3*b* arises as a consequence of the voltage dependence of the number and distribution of the different conductance levels. Figure 3*c* shows the current passing through the channel in the presence of a fivefold KCl gradient. From Fig. 3*c*, we obtained a reversal potential of +35 mV. The diffusion potential expected when cations are completely excluded is +36 mV.

At this point, we would like to discuss the physiological role of K^+ channels in the heart. In Purkinje fibres, Noble and Tsien^{10,19} described three time- and voltage-dependent outward currents in which the main current-carrying species was K^+ . These currents are the pacemaker currents, i_{K_2} (but see ref. 20), and the repolarizing currents i_{K_1} , i_{K_2} . In addition, a fourth time-independent K^+ conductance showing inward-going rectification, i_{K_1} , is also present. More recently, Ca^{2+} -activated K^+ conductances have been shown to contribute to the early repolarization phase of Purkinje action potential²¹ as well as to the steady-state K^+ conductance²². Many of these currents

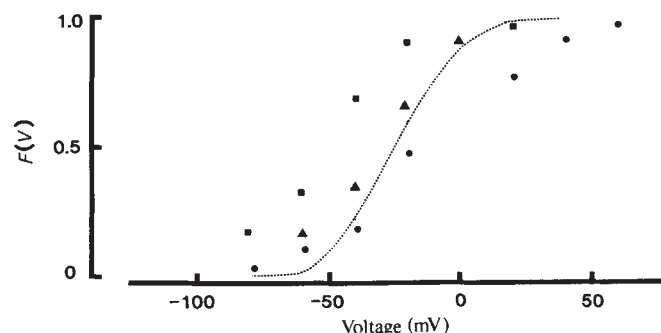


Fig. 4 Single-channel activation curves. The fraction of time in the open state [$F(V)$] of the 28 pS K^+ channel was obtained in three different membranes containing one channel in different ionic conditions. $F(V)$ was calculated from histograms such as that shown in Fig. 2*b*. ■, $F(V)$ in symmetrical 0.1 KCl, 5 mM HEPES-Tris, pH 7.2. ●, $F(V)$ in 1 M KCl, 5 mM HEPES-Tris, pH 7.2. ▲, $F(V)$ in 0.2 M KCl, 5 mM HEPES-Tris, pH 7.2 *cis*, 0.1 M KCl, 5 mM HEPES-Tris, pH 7.2 *trans*. Dashed line is the steady-state conductance voltage curve of the repolarizing current, i_{K_1} , and was taken as described in ref. 2.

are also present in ventricular and atrial fibres^{11,12}. In agreement with the findings in cardiac muscle, we describe here several types of channel in which K^+ is the major current carrier. In particular, Fig. 4 shows that the voltage dependence of the 28 pS K^+ channel is comparable with that of the i_x currents (dashed line) described in ref. 2. Furthermore, a $P_{K^+}/P_{Na^+} = 5$ for this channel correlates well with the finding that i_x currents are not purely K^+ currents¹⁰. Although the assumption that i_x and the 28 pS channel activate with the same polarity of voltage is arbitrary, we are inclined to view the *cis* chamber as equivalent to the cytoplasmic side of the incorporated channel based on the following observations: (1) all 28 pS channels studied gate in the same direction, that is, the direction of Fig. 4; (2) inside-out cardiac sarcolemma vesicles fuse more readily than right-sided vesicles, which in the most simple model of fusion would leave the cytoplasmic face of the incorporated channels in the *cis* side¹⁶; and (3) the same incorporation procedure described here when applied to skeletal muscle membranes results in the incorporation of Ca^{2+} -activated K^+ channels with the Ca^{2+} site on the *cis* side¹⁷. Nevertheless, if the 28 pS channel were, in fact, to be activated in the cell by hyperpolarizing voltages, it would have the gating characteristics of the recently described inward pacemaker current, i_f (ref. 20). Thus, i_f is a time- and voltage-dependent inward current which activates by hyperpolarizing voltages and is a mixed K^+ , Na^+ current²⁰. Chloride ions, on the other hand, although clearly involved in the generation of the cardiac action potential, have not been unambiguously identified with a particular current component^{5,6,13}. This makes it difficult to assign a physiological role to the Cl^- channel found in the present work, so that more experiments need to be performed in both cardiac cells and bilayers to make a complete identification of this channel. Finally, the channels from cardiac sarcolemma can be incorporated into bilayers made from monolayers (data not shown). This method has been recently used to reconstitute the acetylcholine channel²³⁻²⁶. In this type of membrane, we found that the electrical activity is quite complex and indicative of the presence of several types of channel. Using this method, we have been able to identify the 55 pS Cl^- channel and several of the K^+ channels described above. Our choice of the fusion method reported here to incorporate cardiac sarcolemma channels was based on the observation that membranes containing a single channel can be obtained easily. This enormously simplified the analysis of the data.

This work was supported by NIH grants GM-25277 and GM-28992 and by a postdoctoral fellowship to R.C. from the Muscular Dystrophy Association of America. We thank Marianita Sanchez for secretarial help.

Received 22 March; accepted 8 June 1982.

- Noble, D. in *The Initiation of the Heartbeat* 2nd edn, 186 (Clarendon, Oxford, 1979).
- McAllister, R. E., Noble, D. & Tsien, R. W. *J. physiol., Lond.* **251**, 1-59 (1975).
- Rudel, J. & Rudel, R. *Pflügers Arch. ges. Physiol.* **315**, 136-158 (1970).
- Lee, K. S., Weeks, T. A., Kao, R. L., Akaiki, W. & Brown, A. M. *Nature* **278**, 269-271 (1979).
- Carmeliet, E. *J. Physiol., Lond.* **156**, 375-388 (1961).
- Fozzard, H. A. & Hiraoka, M. *J. Physiol., Lond.* **234**, 569-586 (1973).
- Reuter, H. A. *Rev. Physiol.* **41**, 413-424 (1979).
- Colquhoun, D., Neher, E., Reuter, H. & Stevens, C. F. *Nature* **294**, 752-754 (1981).
- DeFelice, L. J. & Clapham, D. E. *Biophys. J.* **37**, 341A (1982).
- Noble, D. & Tsien, R. W. *J. Physiol., Lond.* **200**, 205-231 (1969).
- Beeler, G. W. & Reuter, H. *J. Physiol., Lond.* **268**, 177-210 (1977).
- Brown, H. F., Clark, A. & Noble, S. J. *Physiol., Lond.* **258**, 547-577 (1976).
- Kenyon, J. L. & Gibbons, W. R. *J. gen. Physiol.* **73**, 117-138 (1979).
- Jones, G. R., Maddock, S. W. & Besch, H. R. *J. biol. Chem.* **255**, 9971-9980 (1980).
- Miller, C. & Racker, E. *J. Membrane Biol.* **30**, 285-300 (1976).
- Coronado, R. & Williams, A. *Biophys. J.* **37**, 343A (1982).
- Latorre, R., Vergara, C. & Hidalgo, C. *Proc. natn. Acad. Sci. U.S.A.* **79**, 805-810 (1982).
- Cohen, F. S., Zimmerberg, J. & Finkelstein, A. *J. gen. Physiol.* **75**, 251-270 (1980).
- Noble, D. & Tsien, R. W. *J. Physiol., Lond.* **195**, 185-214 (1968).
- DiFrancesco, D. J. *Physiol., Lond.* **314**, 377-393 (1981).
- Siegelbaum, S. A., Tsien, R. W. & Kass, R. S. *Nature* **269**, 611-613 (1977).
- Isenberg, G. *Pflügers Arch. ges. Physiol.* **371**, 71-76 (1977).
- Schindler, H. & Quast, U. *Proc. natn. Acad. Sci. U.S.A.* **77**, 3052-3056 (1980).
- Nelson, N., Anholt, R., Lindstrom, J. & Montal, M. *Proc. natn. Acad. Sci. U.S.A.* **77**, 3057-3061 (1980).
- Boheim, G. *et al. Proc. natn. Acad. Sci. U.S.A.* **78**, 3586-3590 (1981).
- Labarca, P., Lindstrom, J. & Montal, M. *Biophys. J.* **37**, 170a (1982).

Macrophage binding of *Staphylococcus albus* is blocked by anti I-region alloantibody

John Stewart, Elizabeth J. Glass & Donald M. Weir

Department of Bacteriology, University Medical School, Teviot Place, Edinburgh EH8 9AG, UK

Cell surface interactions involving carbohydrate may be important in immune recognition. Previous work from this laboratory has demonstrated the presence of 'lectin-like' receptors on mouse peritoneal macrophages that bind bacteria by means of their cell wall sugars¹⁻⁴. Others have shown that Ia molecules can bind antigen at specific sites which may be involved in presenting antigen to the immune system⁵ and recent work has shown that these molecules can carry carbohydrate determinants⁶⁻⁷. It has also been found that human Ia molecules can bind to carbohydrates⁸. As cell surface carbohydrate recognition mechanisms have been implicated in other immune interactions⁹⁻¹¹ sugar-specific receptors may have a function in self-non-self recognition. We show here that the binding of the bacterium *Staphylococcus albus* to mouse peritoneal macrophages was inhibited by various conventional and monoclonal antibodies to Ia antigens suggesting that an I-region gene product may be associated with the binding of unopsonized bacteria.

The binding of *S. albus* to glass adherent mouse peritoneal macrophages was measured at 0°C in serum free buffer using a predetermined concentration (approximately 5×10^6 organisms per ml) that gave 25-50% binding to control macrophage preparations⁴. When macrophage monolayers from C3HfBu/Kam mice were pretreated with conventional alloantisera of defined specificity there was dose-dependent inhibition of binding (Fig. 1) as compared to untreated controls, with A.TH anti-A.TL at 1:100, 1:300 and 1:600 ($P < 0.05$) and WS9 at 1:100 ($P < 0.01$) and 1:333 and 1:1,000 ($P < 0.05$). However, A.TL anti-A.TH and WS4 were without effect at all dilutions of antibody used. At a dilution of 1/10 of WS4 (anti D^k) maximal fluorescence was obtained in a direct immunofluorescence test confirming that at the dilution used macrophages were maximally binding this antiserum. Both A.TH anti-A.TL and WS9 react with H-2I^k, as well as H-2S^k and H-2K^k respectively. A.TL anti-A.TH is directed against H-2I^s and S^s while WS4 recognizes H-2I-C^k, S^k and D^k. These results suggest that an I-region determined product is involved

Table 1 Inhibition of binding of *S. albus* to macrophages by antibody H10-2.16 and anti-macrophage antibody

a	H10-2.16	Inhibition of binding (%)
	Unadsorbed	19.9 ± 5.7
	Adsorbed	32.6 ± 7.9
b	Anti-macrophage antibody	
	Dilution	
	$\frac{1}{2}$	2.6 ± 11.2
	$\frac{1}{4}$	1.9 ± 5.9
	$\frac{1}{8}$	-3.1 ± 6.7

a, 1 ml of 1:10 dilution of H10-2.16 antibody incubated with 3×10^9 *S. albus* at 0°C for 2 h. The bacteria were removed by centrifugation and antibody used at a final dilution of 1:50 as described in Fig. 1. Unadsorbed antibody was treated as above but without the addition of *S. albus*. Each figure represents the mean ± 1 s.e. of four experiments. b, The monoclonal antibody specific for macrophages (rat IgG against mouse macrophages. Sera-lab, Sussex) was incubated with macrophage monolayers for 40 min at 0°C before incubation with bacteria as described previously. Each figure represents the mean ± 1 s.e. of three experiments.

in bacterial attachment to macrophages in the absence of opsonins.

In order to further investigate the H-2 region involved, the effect of two monoclonal antibodies on *S. albus* binding to C3HfBu/Kam peritoneal macrophages was determined (Fig. 1). Both antibodies which react with I-A products gave significant inhibition (H 8-109.13; 1:1,000 $P < 0.001$; 1:10⁴, 1:10⁵ and 1:10⁶ $P < 0.01$; Antibody H 10-2.16; 1:20, 1:60, and 1:180 $P < 0.05$), indicating a possible role for an I-A subregion product in bacterial binding. The possibility that the inhibition observed might be due to cross-reactivity of Ia antigens and the bacterial surface is excluded (1) by the extensive

Table 2 The effect of antibody from cell line 8-109.13 on the binding of *S. albus* to macrophages

Dilution	Inhibition of binding (%)		
	C57BL/6	BALB/c	C3H.OH
1:10 ³	3.0±6.1	9.1±11.4	-6.1±5.0
1:10 ⁴	-4.8±11.4	0.9±1.9	1.9±7.6
1:10 ⁵	-2.7±3.1	10.1±5.6	-3.3±5.0
1:10 ⁶	10.0±6.6	-10.1±9.2	6.9±2.2

C57BL/6 (b haplotype) and BALB/c (d haplotype) mice were from the Moredun Institute, Edinburgh, C3H.OH (o2 haplotype) were from Olec 76, Bicester and C3HfBu/Kam were from the departmental breeding colony. Each figure represents the mean ±1.s.e. of three experiments. Female mice age 6-8 weeks were used.

washing of the monolayers prior to the addition of bacteria and (2) by adsorption of the monoclonal antibody H 10-2.16 with *S. albus*. The latter treatment did not significantly alter the inhibitory capacity of the antibody (Table 2). Table 2 also shows that antibody directed against macrophage membrane determinants other than H-2 has no effect on bacterial binding. This indicates that the observed inhibition is a function of the alloantibody used.

The monoclonal antibody from clone 8-109.13 is specific for a private determinant of I-A^k. Therefore this antibody should only inhibit bacterial binding to macrophages from a mouse with k haplotype. The effect of H8-109.13 on the binding of *S. albus* to peritoneal macrophages from various mouse strains, with different haplotypes is shown in Table 3. As described above binding to cells from C3HfBu/Kam mice (k haplotype) was blocked by this antibody (Fig. 1). However, the attachment was unaffected by pretreatment of macrophage monolayers prepared from mice with b, d or o2 haplotypes with this antibody. This observation further suggests that a product of the I-region is involved in unopsonised bacterial recognition and that the observed inhibition by the various alloantibodies was not due to crossreactivity with a general 'lectin-like' receptor showing similar antigenic determinants to a product of I-A^k.

The data do not necessarily imply the 'lectin-like' receptor for bacterial cell wall sugars and the Ia antigen on mouse peritoneal macrophages are identical and there are a number of possible explanations for the findings. For example, Ia antigens may bind bacteria, and possibly other antigens, directly or as postulated by Higgins *et al.*⁷ protein-Ia acting as a glycosyl transferase could catalyse the attachment of bacteria to glycolipid Ia. The observed inhibition of bacterial binding by various alloantibodies could be due to blocking of these structures or inhibition of glycosyl transferase activity. The possibility has also not been excluded that the bacteria bind to a separate receptor which is in close proximity to the I-region determined structure and that the decrease in binding is due to steric hindrance.

Complete inhibition of bacterial binding was never obtained with any of the antibodies used. Similarly, in sugar inhibition studies with glucose or galactose, bacterial binding with *S. albus* to various phagocytes was only partially inhibited by the sugars⁴. These results indicate that the interactions of unopsonised bacteria with macrophages involve a number of recognition processes. These could include binding of bacteria by macrophage 'lectin-like' receptors¹ and attachment mediated by lectin molecules in the bacterial cell wall^{12,13}. It has also been proposed that phagocytes can bind particles by nonspecific mechanisms¹⁴ including hydrophobic interactions.

Results to date in which monolayers were treated with alloantibody and then with glucose or galactose showed no additional inhibition to that seen with alloantibody alone. This suggests that both reagents are reacting with the same site. Further work including cell membrane fractionation procedures is planned to determine the relationship between the macrophage 'lectin-like' receptor previously described by this laboratory and the I-A associated product responsible for the attachment described here.

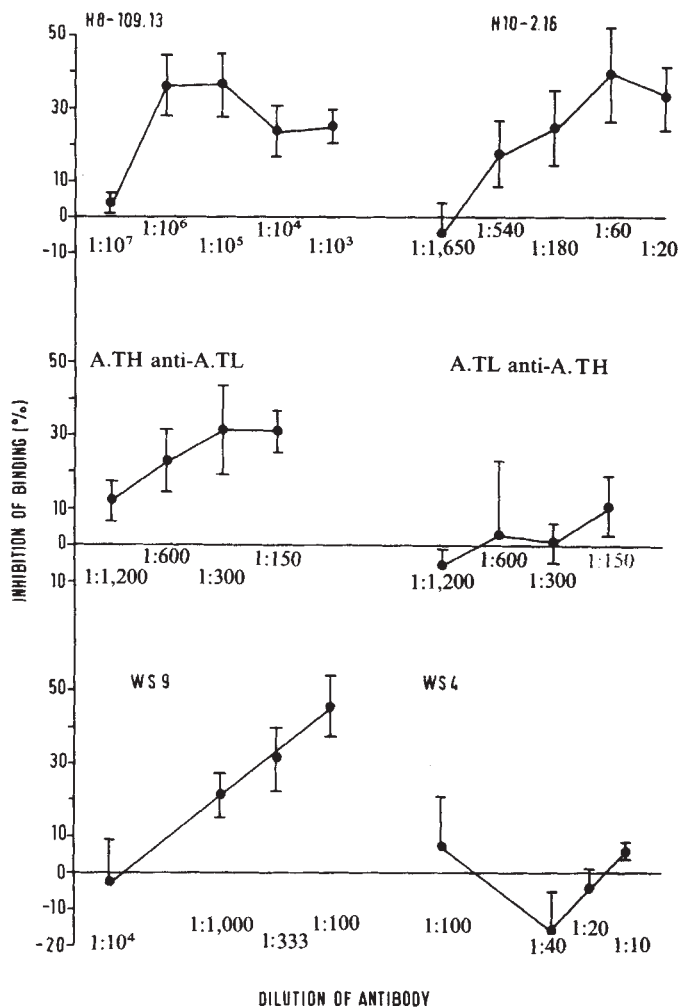


Fig. 1 The effect of various alloantisera on the binding of *S. albus* to C3HfBu/Kam peritoneal macrophages. Each point represents the mean ±1.s.e. of between 3 and 8 experiments. *S. albus* was grown in horse digest broth containing 2.5% glucose, cells were collected in log phase and killed by exposure to 0.5% formalin at 4°C for 24 hours. Mouse peritoneal macrophages, obtained by lavage, were suspended in Eagles minimal essential medium (Wellcome Research Laboratories) without serum, buffered with HEPES and supplemented with penicillin streptomycin and glutamine, at a concentration of 2×10^5 macrophages per ml. One ml of the cell preparation was layered onto 13-mm diameter (No. 1) glass coverslips in 16-mm diameter tissue culture plates (Costar) and incubated for 1 h at 37°C. Non-adherent cells were removed by washing with Dulbecco's phosphate-buffered saline containing Ca^{2+} and Mg^{2+} ions (D.PBS+B). Monolayers were incubated with 0.25 ml of appropriate dilution of alloantibody or diluent control (D.PBS+B) at 0°C for 40 min and then washed twice with D.PBS+B. Coverslips were overlaid with 1 ml *S. albus* in D.PBS+B and incubated for 2 h at 0°C. Cell viability was unaffected during incubations as judged by trypan blue exclusion and general morphology. Non-attached organisms were removed by repeated washing with D.PBS+B and the coverslips air dried, fixed in methanol and stained with May Grünwald/Giemsa. The degree of bacterial binding was estimated microscopically by determining the percentage of macrophage with organisms attached at two or more discrete points. Two hundred macrophages were counted on duplicate coverslips and % inhibition was calculated as the difference between control and test divided by the control $\times 100$.

The I-region of the murine MHC has a major role in regulating the immune system including cell-cell interactions and antigen presentation to lymphocytes. The data described above suggests that macrophages bearing an I-A gene product bind unopsonized bacteria. This interaction may lead to phagocytosis and initiation of the immune response in the non-immune animal.

This work was supported by a grant from the MRC. The monoclonal antibody from cell line 8-109.13, the WS9 and WS4 antisera were a gift from Professor H. Festenstein. The A.TH anti-A.TL and A.TL anti-A.TH were donated by Dr Jana McBride. The monoclonal antibody from cell line 10-2.16 was a gift from Dr E. Unanue.

Received 21 April; accepted 18 June 1982.

- Ögmundsdóttir, H. M. & Weir, D. M. *Clin. exp. Immun.* **26**, 334-339, (1976).
- Freimer, N. B. *et al. Acta path. microbiol. scand. Sect. B*, **86**, 53-57 (1978).
- Ögmundsdóttir, H. M., Weir, D. M. & Marmion, B. P. *Br. J. exp. Path.* **59**, 1-7 (1978).
- Glass, E. J., Stewart, J. & Weir, D. M. *Immunology* **44**, 529-534 (1981).
- Clark, R. B., Chiba, J., Zweig, S. E. & Shevach, E. M. *Nature* **295**, 412-414 (1982).
- Parish, C. R., Higgins, T. J. & McKenzie, I. F. C. *Immunogenetics*, **6**, 343-354 (1978).
- Higgins, T. J., Parish, C. R., Hogarth, P. M., McKenzie, I. F. C. & Hämmerling, G. J. *Immunogenetics* **11**, 467-482 (1980).
- Mann, D. L. & Muchmore, A. J. *Immun.* **124**, 2218-2221 (1980).
- Amsden, A., Evans, V., Yoshida, T. & Cohen, S. J. *Immun.* **120**, 542-549 (1978).
- Fischer, A., Ballet, J. & Criselli, C. *J. clin. Invest.* **62**, 1005-1013 (1978).
- Muchmore, A. & Blaese, R. M. in *Proc. Brooke Lodge Symposium* 505-517 (Academic, New York, 1979).
- Bagg, J., Paxton, I. R. & Weir, D. M. *J. clin. Lab. Immun.* **6**, 165-168 (1981).
- Eshdat, Y., Ofek, I., Yahsour Gan, Y., Sharon, N. & Mirelman, D. *Biochem. biophys. Res. Commun.* **84**, 1551-1559 (1978).
- Capo, C., Bongrand, P., Benoliel, A. M. & Depieds, R. *Immunology* **36**, 501-608 (1980).

A molecular basis for the two locus model of human complement component C4

M. H. Roos*, E. Mollenhauer†, P. Démant* & C. Rittner†

* Division of Genetics, Antoni van Leeuwenhoekhuis, The Netherlands Cancer Institute, Amsterdam, The Netherlands

† Institute of Legal Medicine, University of Bonn, D-5300 Bonn, FRG

The major histocompatibility complex(MHC)-linked fourth component of complement (C4) shows a high degree of polymorphism in several animal species¹⁻⁴. In man C4 polymorphism was detected by distinct charge differences of the variants⁵⁻⁷. O'Neill *et al.*⁸ showed that this C4 polymorphism was controlled by two closely linked genetic loci, *F* (C4A) and *S* (C4B) and these results were extended by Awdeh *et al.*⁹ with an improved typing method. Biochemical analysis of human C4 has revealed that it consists of three polypeptide chains, α , β and γ ¹⁰. In all reports so far on the molecular analysis of human C4¹⁰⁻¹³, no molecular weight differences between the *A* and *B* locus-encoded molecules have been noticed. Here we demonstrate that the C4A and C4B locus-encoded α -chains have a molecular weight (MW) of 96,000 and 94,000, respectively, presenting for the first time a molecular basis for the difference between all C4A and C4B variants tested. Even rare variants that are difficult to allocate to the *A* or *B* locus on the basis of charge differences^{14,15} could be identified as C4A or C4B variants in this way, thereby providing new insights into the relationships between the C4A and C4B loci.

Previous analysis of human C4 by SDS-polyacrylamide gel electrophoresis (SDS-PAGE) showed that the α -, β -, and γ -chains have a MW of about 95,000, 75,000 and 30,000 respectively for C4 from both loci¹⁰⁻¹³. By modifying the ratio of acrylamide to *N,N'*-methylene bisacrylamide (bis) in the gels, the separation properties of the gels are changed since both size and shape of the molecule determine the electrophoretic mobility in the gel. Thus, when C4 from several phenotypes, typed by agarose electrophoresis, was analysed by modified

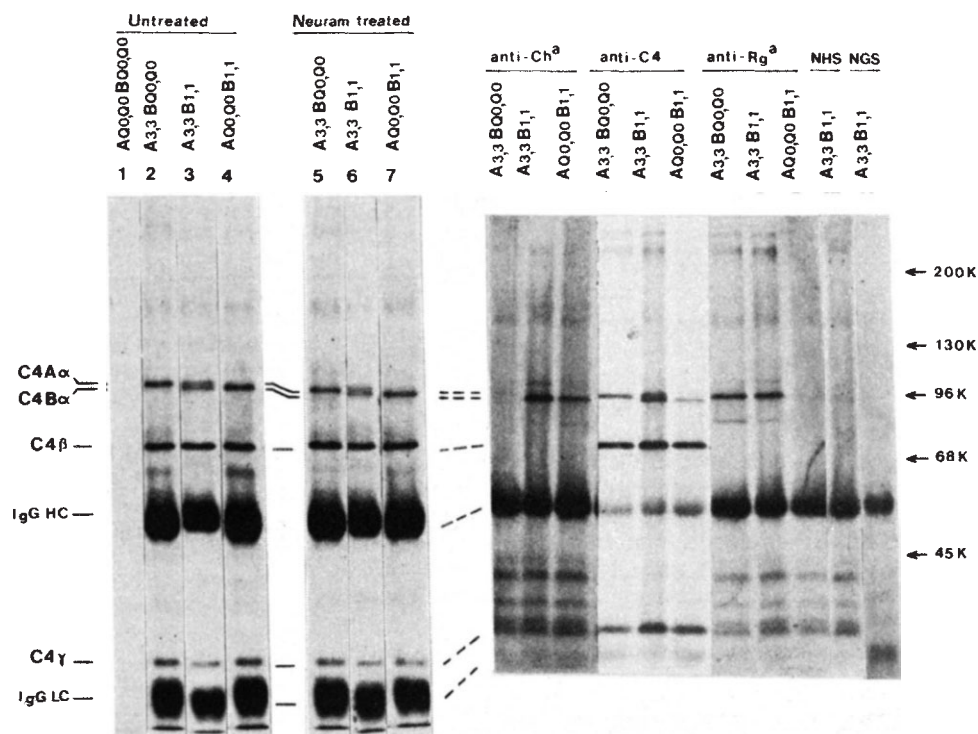
SDS-PAGE, a difference in MW between the α -chains encoded by the *A* and *B* loci became visible (Fig. 1). Plasma C4 phenotypes were determined by agarose electrophoresis according to the method of Awdeh *et al.*⁹, designating the C4A variants A1, A2 and so on, the C4B variants B1, B2 and so on, and silent alleles *QO*. Figure 1a shows the SDS-PAGE analysis of the most common variants for each locus, A3 and B1, present in three phenotypes A3,3BQ0,Q0, A3,3B1,1,AQ0,Q0,B1,1 and as a negative control AQ0,Q0BQ0Q0 (lanes 1-4). The α -chain of the A3 variant has a MW of 96,000, while the α -chain of the B1 variant has a MW of 94,000. No difference in the β - and γ -chains was visible. As agarose typing of the C4 variants is usually carried out after removal of sialic acid with neuraminidase⁹ we also analysed the same phenotypes after desialization (see Fig. 1a, lanes 5-7). A similar reduction of ~3,000 in the MW of both the C4A and C4B α -chains was visible. No change in MW of the β - and γ -chains was found.

To confirm the difference in MW between *A* and *B* locus-encoded α -chains, anti-Rodgers (Rg^a) and anti-Chido (Ch^a) sera, which have a known specificity for C4A and C4B molecules respectively¹⁶⁻¹⁸ were used. C4, precipitated from neuraminidase-treated plasma of different haplotypes, was analysed by SDS-PAGE, followed by blotting of the proteins in the gel onto nitrocellulose sheets as described previously¹⁹. The sheets were then reacted with anti-Ch^a, anti-C4 or anti-Rg^a; the results are shown in Fig. 1b. The three C4 phenotypes reacted with anti-C4, revealing the C4 α -, β - and γ -chains (lanes 4-6), in addition to some background bands (lane 11). With anti-Ch^a only the C4B α -chains became visible (Fig. 1b, lanes 1-3), and with anti-Rg^a only the C4A α -chains were detected (lanes 7-9) over the background (lane 10). Samples which had not been treated with neuraminidase gave qualitatively similar results. Also, when unreduced C4 was analysed by SDS-PAGE in a 6% gel, a MW difference was found between native C4 from the *A* and *B* loci. After blotting and analysis of the variants A3 and B1 with anti-Ch^a and anti-Rg^a a reaction of the C4A molecule with anti-Rg^a, and a reaction of the C4B molecule with anti-Ch^a was found (data not shown).

To establish whether the difference in α -chains between C4A and C4B variants applies also to other variants than the most common ones, we have analysed several different phenotypes, including some with rare and new C4 variants. The agarose electrophoresis patterns of 14 phenotypes are shown in Fig. 2a; as the pattern is highly complex a schematic presentation of the individual C4A and C4B variants is shown in Fig. 3. The C4A variants show generally a higher mobility than the C4B variants but some variants do overlap. This is especially the case for the B4, B5 and B5.1 variants which do overlap with most C4A variants and for the A1 variant which overlaps with most C4B variants. In this case C4A or C4B variants can only be identified by genetic studies and by analysing the functional activity with a lytic overlay (Fig. 2b), since the functional activity of the C4B variants is about 10-fold higher than C4A variants^{9,14}. As can be seen from Fig. 2a and b the A1 variant (lane 5), like other C4A variants, does not show lytic activity, while the B4, B5 and B5.1 variants show a lytic activity similar to all other C4B variants. When the different phenotypes were analysed by SDS-PAGE (Fig. 4) all the C4A variants A1, A2, A3, A4, A5.1 and A6 (A5.1 not shown), showed an α -chain with a MW of 96,000 while all the C4B variants, B-2, B1, B1.1, B2, B4, B5 and B5.1 (B-2 and B4 not shown) showed α -chains with a MW of 94,000. Thus with this method even variants that overlap in the agarose gel electrophoresis can be identified structurally as C4A or C4B variants.

As some informative phenotypes were not available, however, our experiments do not exclude the possibility that the A4 and A5.1 variants contain an α -chain with a 94,000 MW in addition to the 96,000 α -chain; nor can it be excluded that the B-2, B1.1, B2, B4, B5 and B5.1 variants contain a 96,000 MW α -chain in addition to the 94,000 α -chain. This is, however, unlikely as the variants A1, A3, A6 and B1, which

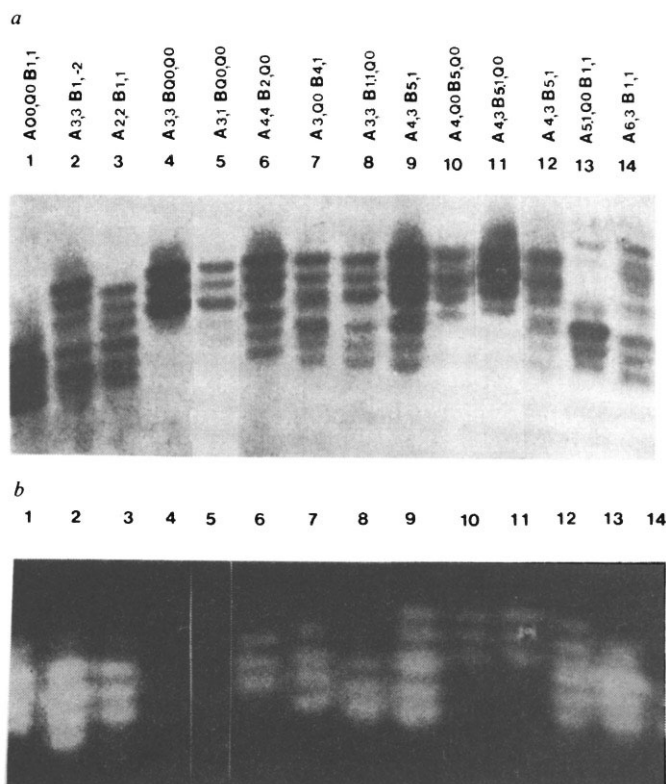
Fig. 1 Polymorphism of the C4 α -chains as detected by SDS-PAGE of human C4 and neuraminidase-treated human C4. *a*, C4 was immunoprecipitated from human EDTA-plasma with goat anti-human C4 (Atlantic Antibodies). The precipitates were washed, disintegrated and reduced as described previously²⁰, except that 10% 2-mercaptoethanol was used. SDS-PAGE was carried out according to a modification of the method of Laemmli *et al.*²⁸; a 10% acrylamide gel was used with a ratio acrylamide: *N,N'*-methylene bisacrylamide (bis) of 1:0.006. For the separation of unreduced C4, a 6% acrylamide gel was used with an acrylamide: bis ratio of 6:0.0013. To prevent the gel from sticking to the glass plates, 0.3% agarose (Ultra Pure; Bio-Rad) was incorporated in the gel. *b*, After electrophoresis, proteins in the gel were blotted on to nitrocellulose sheets by compressing the gel for 1 h between two of these sheets and a 2-cm layer of tissues¹⁹. Thereafter, the gel was stained with Coomassie blue R250. The nitrocellulose sheets were incubated overnight with 15 ml blotting buffer (500 mM NaCl, 10 mM sodium phosphate (pH 7.3), 10 mM EDTA, 0.25% gelatin, 0.5% Triton X-100) supplemented with 50 μ l anti-Ch^a serum, 20 μ l anti-C4 serum or 50 μ l anti-Rg^a serum. Then the sheets were washed for 1 h with 50 ml blotting buffer, incubated for 1 h with horseradish peroxidase-conjugated rabbit antiserum prepared against human IgG or goat IgG (DAKO). The sheets were washed for 1 h before a fresh solution containing 0.5 mg ml⁻¹ phosphate-buffered saline (PBS) of the peroxidase substrate, 3,3'-diaminobenzidinetetrahydrochloride and 0.3% H₂O₂. The reaction was stopped after 10 min by washing with PBS. Neuraminidase treatment of human plasma was carried out according to Awdeh *et al.*⁹. *a*, SDS-PAGE: lanes 1–4, C4 precipitated from EDTA-plasma; lanes 5–7, C4 precipitated from neuraminidase-treated EDTA-plasma. *b*, Blots treated with: lanes 1–3, anti-Chido serum; lanes 4–6, anti-C4 serum; lanes 7–9, anti-Rodgers serum; 10, normal human serum (NHS); 11, normal goat serum (NGS).



were analysed in combination with 2 *QO* alleles present at the other locus, only showed one α -chain. In all cases where one or two C4A variants were present, an α -chain with a MW of 96,000 was found whereas an α -chain with a MW of 94,000 was always found when one or two C4B variants were present. Since the typing of the variants in agarose gels (Fig. 2) was carried out with neuraminidase-treated samples we analysed

all the phenotypes described above by SDS-PAGE after neuraminidase treatment. This resulted, as shown in Fig. 1 for the A3 and B1 variants, in a decrease in the apparent MW of 3,000 for both the C4A and C4B α -chains; the difference between the α -chains of the C4A and C4B variants remained (data not shown). When comparing the concentration in the gel of the C4A α -chains with the C4B α -chains (Fig. 4)

Fig. 2 Human C4 polymorphism as detected by agarose gel electrophoresis. EDTA-plasma samples were treated with neuraminidase (5 U per μ l plasma; type VIII, Sigma) as described by Awdeh *et al.*⁹. Agarose gel electrophoresis was carried out according to the method of O'Neill *et al.*⁸, on 20 $\times$ 20 cm glass plates covered with 2 mm 0.5% agarose (Litex HSA), for 5 h using a voltage gradient of 20 V cm⁻¹ until the haemoglobin marker had migrated at least 10 cm from the origin (some variants may only be recognized in such extended runs); two plates were run in parallel. After completion, one plate was overlaid with goat anti-human C4 serum (Atlantic Antibodies) overnight, washed, dried and stained with Coomassie blue. The duplicate plate was covered with an agar overlay containing a haemolytic system with C4-deficient guinea pig serum as described previously²⁹. The C4 variants were attributed to the C4B locus when they caused a lytic banding pattern in the functional overlay^{9,14}. The individual C4A and C4B locus variants were identified by comparing the immunofixation banding patterns and lytic banding patterns with a reference set distributed by Dr G. Mauff to C4 typing laboratories in 1979 and the final designation of human C4A and B allotypes (A1, A2, ... B1, B2 ...) is in accordance with Awdeh *et al.*⁹ and with a proposal of Drs G. Hauptmann, G. Mauff and C.R. (see also Fig. 3). The influence of silent alleles (*QO*) on the amount of C4A and C4B molecules present in the phenotypes was judged by comparing the relative strength of the A and B locus-encoded variants. In addition, in some cases, HLA-A, B, C, DR information (linkage disequilibrium) was also considered. Some variants could be followed in families by segregation¹⁵. *a*, Immunofixation banding pattern of 14 C4 phenotypes; *b*, haemolytic overlay banding pattern of the same samples.



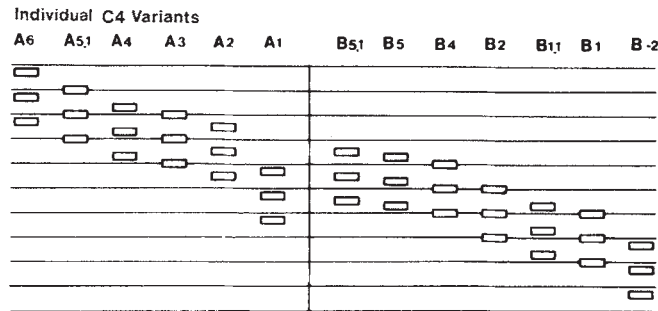


Fig. 3 Schematic representation of the C4 variants used in this study. These are all the variants that could be identified at present in the Bonn laboratory. Note the designation of the variants A5.1, B1.1, B5.1 and B-2.

sometimes a clear difference in the intensity of the α -chains was observed. Whether this represents the presence of a silent allele at the locus with the fainter band is under investigation¹⁵.

Our results constitute the first evidence for a molecular difference between all C4A and C4B variants tested. All the C4A variants analysed contain an α -chain with an identical MW of 96,000 whereas all the analysed C4B variants contain an α -chain with an identical MW of 94,000. This indicates that the C4A and C4B variants differ structurally from each other even though some variants among both loci have very similar mobilities on agarose electrophoresis. The difference between C4A and C4B α -chains is another indication that the C4 structural genes are located in the *HLA* complex, and also provides a new analogy between human and mouse C4. In the mouse a different MW for the α -chains of C4 and the allotype Slp (sex limited protein) has been described^{20,21} providing evidence that distinct structural C4 and Slp loci are linked to the S region of the H-2 complex. Another striking similarity between human

and mouse C4 is the difference in functional activity between the products of the two loci. Human C4B variants possess a functional activity that is 10 times higher than that of the C4A variants^{9,14}; in the mouse the C4 locus produces a functionally active C4 molecule, while the Slp locus produces a molecule for which no functional C4 activity has been detected^{21,22}. This could be a consequence of the structural difference between the α -chains of the two loci since the α -chains have been shown to be involved in the activation of C4²³.

The antisera directed against the *HLA*-linked Chido and Rodgers blood group antigens^{24,25} were shown previously to react specifically with the α -chains of the C4B and C4A variants, respectively, by the agglutination test and by inhibition of the antiglobulin test^{17,18}. Our blotting method proved to be a new and sensitive way of analysing the reaction of the anti-Ch^a and anti-Rg^a sera with C4 α -chains of the A3 and B1 variants. Thus C4 variants, so far not found in a haplotype with a QO allele at the other C4 locus and therefore considered to be Rodgers- or Chido-positive only by analogy or inference, can be analysed directly with our method for their reaction with anti-Rg^a or anti-Ch^a serum.

The removal of sialic acid from native C4 by neuraminidase affects the α -chains, as expected¹². This result, together with the localization of both the Rg^a and Ch^a antigenic determinants^{17,18} and the charge difference between the human C4 variants on the C4 α -chain²⁶, indicates that parts of the α -chain are exposed on the C4 molecule. However, the human C4 β - and γ -chains are not affected by neuraminidase, although the human C4 β -chain, in contrast to the murine C4 β -chain, also contained sialic acid^{12,27} implying that at least the carbohydrate part of the β -chain is hidden inside the C4 molecule.

Our findings may open new ways for the biochemical characterization of the polymorphism of human C4 and of the genetic organization and structure of the human C4 genes. In addition, our results extend the homologies between *HLA* and *H-2* and may provide additional possibilities for investigation and understanding of the association of the early complement component genes with the MHC.

We thank Dr T. Meo for a gift of the C4-deficient serum and Drs C. Giles and R. Berger for the anti-Chido and anti-Rodgers sera and Mrs N. van Nuland-van Wijngaarde for secretarial assistance. This study was supported by the Queen Wilhelmina Foundation (Netherlands Cancer Foundation) and by the Deutsche Forschungsgemeinschaft (Ri 164-15-2).

Received 4 May; accepted 5 July 1982.

- Rittner, C. *Hum. Genet.* **35**, 1-20 (1976).
- Shreffler, D. C. *Transplant Rev.* **32**, 140-167 (1976).
- Shreffler, D. C., Atkinson, J. P., Brown, L. J., Parker, K. L. & Roos, M. H. in *7th International Convocation on Immunology* Niagara Falls, New York, 1980 (eds Zaleski, M., Abeyounis, C. J. & Kano, K.) 78-88 (Karger, Basel, 1981).
- Alper, Ch. A. in *The Role of the Major Histocompatibility Complex in Immunobiology* (ed. Dorf, M. E.) 173-220 (Garland STPM, New York, 1981).
- Teisberg, P., Åkeson, P., Olaisen, B., Gedde Dahl, T. Jr & Thorsby, E. *Nature* **264**, 253-254 (1976).
- Mauff, G., Bender, K. & Fisher, B. *Vox Sang* **34**, 296-301 (1978).
- Rittner, C. *et al. Immunobiology* **158**, 119-128 (1980).
- O'Neill, G. J., Young Yang, S. & Dupont, B. *Proc. natn. Acad. Sci. U.S.A.* **75**, 5165-5169 (1978).
- Awdeh, Z. H. & Alper, C. A. *Proc. natn. Acad. Sci. U.S.A.* **77**, 3567-3580 (1980).
- Schreiber, R. D. & Müller-Eberhardt, H. J. *J. exp. Med.* **140**, 1324-1335 (1974).
- Bolotin, C., Morris, S., Tack, B. & Prah, J. *Biochemistry* **16**, 2008-2015 (1977).
- Gigli, I., Von Zabern, I. & Porter, R. P. *Biochem. J.* **165**, 439-446 (1977).
- Lundwall, A., Malmheden, I., Stalenheim, G. & Sjöquist, J. *Eur. J. Biochem.* **117**, 141-146 (1981).
- Mauff, G. *Forensic Sci. Int.* **18**, 291 (1981).
- Rittner, C. *et al.* (in preparation).
- O'Neill, G. J., Yang, S. Y., Tegoli, J., Berger, R. & Dupont, B. *Nature* **273**, 668-679 (1978).
- Tilley, C. A., Romans, D. G. & Crookston, M. C. *Nature* **276**, 713-715 (1978).
- Chu, V. F. H., Marsh, W. L. & Gigli, I. *J. Immun.* **128**, 181-185 (1982).
- Roos, M. H. & Démant, P. *Immunogenetics* **15**, 23-30 (1982).
- Roos, M. H., Atkinson, J. P. & Shreffler, D. C. *J. Immun.* **121**, 1106-1115 (1978).
- Ferreira, A., Nussenzweig, A. & Gigli, I. *J. exp. Med.* **148**, 1186-1197 (1978).
- Carroll, M. C. & Capra, J. D. *Proc. natn. Acad. Sci. U.S.A.* **75**, 2424-2428 (1978).
- Müller-Eberhardt, H. J. *A. Rev. Biochem.* **44**, 697-724 (1975).
- Midleton, J. *et al. Tissue Antigens* **4**, 366-373 (1974).
- Giles, C. M. *et al. Tissue Antigens* **8**, 143-149 (1976).
- Merag, B., Olaisen, B. & Teisberg, P. *Scand. J. Immun.* **14**, 303-307 (1981).
- Roos, M. H., Kornfeld, S. & Shreffler, D. C. *J. Immun.* **124**, 2860-2863 (1980).
- Laemmli, U. K. *Nature* **227**, 680-682 (1970).
- Rittner, C. & Mollenhauer, E. *Ärzt. Lab.* **27**, 232-238 (1981).

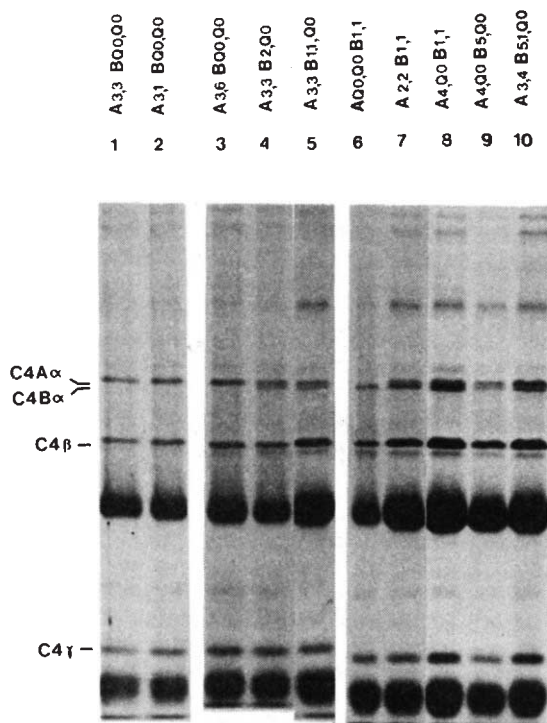


Fig. 4 Difference in MW of the C4A and C4B locus-encoded α -chains in 10 different phenotypes. Immunoprecipitation, SDS-PAGE and staining with Coomassie blue was performed as described in Fig. 1 legend. Lanes 1-10, analysis of 10 different phenotypes; lanes 3, 4, 5 are from a different gel. The total amount of C4 present in the samples in the gel is no indication of the level of C4 present in the plasma, as the C4 in the samples was adjusted to give clearly visible bands.

Nucleotide sequence of the rat skeletal muscle actin gene

R. Zakut, M. Shani, D. Givol*, S. Neuman, D. Yaffe & U. Nudel

Departments of Cell Biology and *Chemical Immunology, The Weizmann Institute of Science, Rehovot 76100, Israel

The actins constitute a family of highly conserved proteins found in all eukaryotic cells. Their conservation through a very wide range of taxonomic groups and the existence of tissue-specific isoforms make the actin genes very interesting for the study of the evolution of genes and their controlling elements. On the basis of amino acid sequence data, at least six different mammalian actins have been identified (skeletal muscle, cardiac muscle, two smooth muscle actins and the cytoplasmic β - and γ -actins)¹⁻⁵. Rat spleen DNA digested by the *Eco*RI restriction enzyme contains at least 12 different fragments with actin-like sequences but only one which hybridized, in very stringent conditions, with the skeletal muscle cloned cDNA probe⁶. Here we describe the sequence of the actin gene in that fragment. The nucleotide sequence codes for two amino acids, Met-Cys, preceding the known N-terminal Asp of the mature protein. There are five small introns in the coding region and a large intron in the 5'-untranslated region. Comparison of the structure of the rat skeletal muscle actin gene with the available data on actin genes from other organisms shows that while the sequenced actin genes from *Drosophila* and yeast have introns at different locations, introns located at codons specifying amino acids 41, 121, 204 and 267 have been preserved at least from the echinoderm to the vertebrates. A similar analysis has been done by Davidson⁷. An intron at codon 150 is common to a plant actin gene and the skeletal muscle acting gene.

Act15 is a recombinant bacteriophage containing the coding region of the rat skeletal muscle actin gene and 11 kilobases (kb) of its 5'-flanking region⁶. A 3.7-kb *Eco*RI DNA fragment was subcloned in pBR322 (pAct15-2); the restriction map of the subclone and the strategy used for sequencing are shown in Fig. 1. The sequence of the entire coding region and the 3'-untranslated region is given in Fig. 2a.

The coding sequence begins with codons for two amino acids, methionine and cysteine, which are absent from the mature protein. They are followed by the codon for Asp (GAC), the known N-terminal residue of actin. It is of interest that in all six *Drosophila* actin genes, the same two codons precede the codon specifying the N-terminal amino acids⁸. The translation initiator codon ATG was found in all eukaryotic genes. The

preservation of the cysteine codon in actin genes of such remote organisms suggests a functional role for this sequence. The primary mRNA translation product probably contains the Met-Cys residues, which are then cleaved off the polypeptide chain. Note, however, that the Cys codon is absent from the rat β -actin gene (U.N. *et al.*, unpublished).

The amino acid sequence deduced from the nucleotide sequence is identical to that of rabbit skeletal muscle actin. (The sequence of rat skeletal muscle actin has not been determined.)

A recombinant cDNA plasmid (p106) containing the entire nucleotide sequence of the 3'-untranslated region of the rat muscle actin mRNA was isolated and sequenced^{9,10}. Comparison of the 3' region of the Act15 insert with p106 showed that the Act15 insert contains 75 base pairs (bp) of the 3'-untranslated region of the structural gene and lacks the last 165 bp of the 3' end. The only differences in the overlapping region were two additional thymines in p106 and one A/T replacement. These differences may have arisen from mistakes in the sequence of p106 or from polymorphism in the sequence of the 3'-noncoding region of this gene. As the sequence of the terminal 165 nucleotides of the 3'-untranslated region was deduced from a cDNA clone, the existence of an intron in that region cannot be excluded.

The skeletal muscle actin gene contains five introns in the coding region (Fig. 1c). The sequences of the splice sites are in accordance with the consensus sequence for splice sites suggested by Breathnach and Chambon¹¹ (Fig. 2a). At the codon specifying amino acid 150, the intron sequence shows another potential splice site (CAG/GTA), which is present 32 bp downstream from the CG/GT splice site at codon 150 (Fig. 2a). The sequence between the two splice sites encodes 11 amino acids. Thus, if this site were involved in mRNA splicing, it could add an additional peptide (Gly-Glu-Cys-Ser-His-Thr-Ala-His-Ser-Gly-Pro) after Thr 149, without changing the reading frame downstream. This would produce a protein having the exact sequence of skeletal muscle actin, with an insert of 11 amino acids. The existence of such a protein has not been reported so far.

We have previously demonstrated by electron microscopy that the rat skeletal muscle actin gene has a large intron, ~750 bp long, at its 5' end⁶. The sequencing data (Fig. 2a) do not reveal such a large intron in the coding region, therefore it should be localized in the 5'-untranslated region. The 5' and 3' borders of this intron were determined by S₁ endonuclease mapping¹², and were assigned to nucleotides 250/251 and nucleotides 1,206/1,207, respectively (that is, the intron is located 12 nucleotides upstream from the initiator ATG). The sequence of the two splice sites is in agreement with the consensus splice site sequences¹¹ (Fig. 2b). The sequence data indicate the existence of a potential Goldberg-Hogness box, TATAAA¹³, starting at nucleotide 161, and a potential

Table 1 Positions of introns in actin genes

Source and type of actin gene		Intron location*									
<i>Dictyostelium</i> (several genes) ¹⁸		None									
Yeast ^{19,20}		4	—	—	—	—	—	—	—	—	—
<i>Drosophila</i> ⁷	<i>DmA</i> 6	—	—	—	—	—	—	—	—	307	—
	<i>DmA</i> 4	13	—	—	—	—	—	—	—	—	—
	<i>DmA</i> 2	5'-UT (-8)	—	—	—	—	—	—	—	—	—
Sea urchin gene J†		—	—	41	121	—	203	267	—	—	—
Chicken skeletal muscle‡		5'-UT (-12)	—	—	41	—	150	204	267	—	327
Rat skeletal muscle		5'-UT (-12)	—	—	41	—	150	204	267	—	327
Rat β -cytoplasmic§		5'-UT (-6)	—	—	41	121	—	—	267	—	327
Soybean ¹⁶		—	20	—	—	—	150	—	—	—	355

* The numbers indicate the codons at which introns are localized. Numbers in parentheses indicate the position of introns in the 5'-untranslated (5'-UT) region (the number of nucleotides upstream from the initiator ATG). No introns were found in the sites indicated by dashes.

† W. R. Crain and A. D. Cooper, personal communication.

‡ C. Ordahl, personal communication.

§ Ref. 6 and U.N. *et al.*, unpublished.

11) has indicated that whereas the nucleotide sequence of introns has changed rapidly during evolution, the location of introns in relation to the coding sequence has been highly conserved. On the other hand, the data on the actin genes offer a comparison of gene structure across a much greater evolutionary range, that is, from unicellular organisms to fungi, insects and vertebrates. From the available data, it is apparent that

20 40 60 80
ACCTGGCCAGGGCCAGTCTCTCTCTTGTGTGTCAGGAGACCCGGGCGGAGCTGAGAAGCAGCCGAAG

100 120 140 160
GGACTCTAGTGGCCCAACCAATATGGCTTGGGAAGGSCAACACATTCCTTCGGGCGGTGTGGAGCTCAGGACTA

180 200 220 240
TATAAAACCTGAGGCTAGGAGCAGGGGTACACGGACGTGAAGCTCACTTCCTCACTGGCACCAGGCCAGAGCT

260 280 300 320
AGAGCAGCAGCTAGGCTGGAGGTGGGAGGCTGACCTGGAGACCCAGCAGAGAACTATTGAGCCTTGGGCGTGTATT

340 360 380 400
TGTCACTGGCTGTGGAATTTCTCCAAACTCATATCCAGCCCATTTGGGATGCTCACTTTTGTCTGGTGGAGGGGTG

420 440 460 480
TCTGACACCATCATCTTCCACCTCAGACTCTACCTCCGCCCCCCCCCTGCTGGGAGCTAACTCGAGTCTCTCTCGAA

500 520 540 560
GTGAGCAGCAAGCTGTTTCTACCGAGCCACACACAGCCCTCGTGTGGTGGAGATTTCTCTCCCTGAATTAAGTTCCCC

580 600 620 640
TCTTGCCCTTCACTCCCTCTCTCTCCCCACCTCTCTTCTCTGGGATCAAAATCTGGGCCCTGTGTATGCAGAGGTGG

660 680 700 720
GCTGGATCTCCACAGAGCCAGCCCTGGGAGACTGACTTGAAGATGAGTTTGGAGAGAGGCAGCAATGTCTTAGCCTT

740 760 780 800
CTGAGCTCAAGGCTTTCTAAAGCCCAAGCACCGTGCCAGGCTACTTCTCTGAATAGGAGGGGGGTGGAAGGTTCC

820 840 860 880
TCTTGAGTTGGATGCCGTGGGGCTCTCAAACGGGCAATCTCTTCCAGAGTTGGGACAGATTAACCTCTCTTGGGCAAAA

900 920 940 960
TCTTGCCTGCTCATAGCACTGCCAGGTGACCTTAATGCTCTCTTACAGATTTGAGATTTGAGTAGAGAGGAATAGCTGC

980 1000 1020
GCGCTAGCTGTGACATGTGTGGCCACACAGCGAAGGCTGAGGATCTCAGGATGTCTCTGCATGCCACTATGGAGGAGTAT

1040 1060 1080 1100
AAGTGCCCAACAACCTTTCTCCATAGCAAAAAGCACAGGCTTGACCTGGGAGAGCATCTCTGGCCCTGTGCAGAGGAGG

1120 1140 1160 1180
GCTGTGGGGGAGGAGGTGTGTGGATCTCCAAAGTGAAGTGGGTAAAGCAGAGCTTCACCTTAACCCCATCACATCTC

1200 Met Cys Asp Glu Asp Glu Thr
ACAACGCTCAACACTTCTCTTCTGGAG/AAATCAGACACC ATG TGT GAC GAG GAC GAG ACC

G/GCTGGCTCCGGCTTCGCTGCTTGGCTCGGGAATCAATTTTCCCTCATCCCAAGCCTCAGACTACGGGGCTCTCTCCCGCGGGTGCCCTTAG
270
TCTGACTCTGGTGTCTACACTGCCCTCTCCCGCGGACACAG/GT ATC GAG TCT GCG GGG ATC CAT GAG AOC ACC TAC
280 300
Asn Ser Ile Met Lys Cys Asp Ile Asp Ile Lys Asp Lys Asp Lys Cys Ile His Glu Thr Thr Thr
AAC AGC ATC ATC AAG TGC GAC ATC GAC ATC AGG AAG GAC TGC TAC GCC AAC AAN Val Met Ser Gly Gly
310 320
Thr Thr Met Tyr Pro Gly Ile Ala Asp Arg Met Gln Lys Glu Ile Thr Ala Leu Ala Pro Ser Thr Met
ACT ACC ATG TAC CCC GGT ATC GCT GAC CGC ATG CAC AAG GAG ATC ACA GCT CTG GCT CCC AGC ACC ATG

Lys Ile Lys
AAG ATC AAG/GTGGATGACGGCGCTGGTACGGGGCGGAGACAGCGGGCGGGGAACTCGCAGCCACTGCACACTCTGTCTTCCAG/
340
Ile Ala Lys Pro Pro Glu Arg Lys Tyr Ser Val Thr Ile Gly Gly Ser Ile Leu Ser Leu Ser Thr
ATC ATC ATC GCC CCT CCG GAG GGC AAG TAC TCA GTT TGC ATC GGC GGC TCC ATC CTG GCC TCG TCC TCC
360 370
Phe Gln Gln Met Trp Ile Thr Lys Gln Glu Tyr Asp Glu Ala Gly Pro Ser Ile Val His Arg Lys Cys
TTC CAC CAG ATG TGC ATC ACC AAG CAG GAG TAC GAC GAG GCC GGC CCG TCC ATT GTG CAC GGC AAA TGC

Phe Ter 20 40 60 80
TTC TAC GCGCAACCGGCTCTGTGTGCGGCTCTCTCTCTCAGGACGACAATCGACCATCGTGCATGGTGCAGGGTGCGCCATCTCTC
100 120 140 160
CGCGGTGGCTCATCGGCCCATCGACCGCGGCGCTGTTTGAAGTGTACATAGATTGCTGCTGTTTACCTATTGTGTTATTTTCAA
180 200 220
CAAGCGCTCTGGAAGGAATAGAAACTTGAAGCATTAAGGACGAGCCATCTTTTCTGCTC

despite the great conservation of the amino acid sequence of actins, the number of introns and their positions vary greatly (Table 1).

Clearly, there is no similarity in the number and location of introns of actin genes between fungi, protostomes and the known actin genes of deuterostomes (except for the possible homology regarding the intron in the 5'-untranslated region of rat α - and β -actin genes and the *Drosophila* actin gene *DmA* 2)⁸. However, within the deuterostomes, all known intron positions in sea urchin actin genes (at codons 41, 121, 203 and 267) are found in the rat skeletal muscle actin gene or in the cytoplasmic actin gene, or in both. In addition, the vertebrate actin genes do contain introns which are not found in the known sea urchin actin genes (at codon 327 in α - and β -actin genes and at codon 150 in the α -actin gene). The locations of introns in the chicken muscle actin gene are identical to those in the rat muscle actin gene.

The great resemblances between the actins exclude the possibility of convergent evolution from different ancestral genes. It can be concluded, therefore, that the intron locations, although highly conserved when examined on a small evolutionary scale (that is, among the vertebrates), do change during evolution.

Despite the incomplete data, the homology of intron sites may be used as an indication of the evolutionary relatedness of actin genes. It seems that differences between intron locations in actin isogenes can be regarded as an indication of an ancient separation of these genes. Thus, the great differences in intron locations among the *Drosophila* actin genes suggest that these genes separated from one another very early in evolution. On the other hand, the available data (Table 1) on the intron locations in deuterostomes indicate that these actin genes evolved from a common ancestor more recently. Although all *Drosophila* actin genes apparently have been isolated and partially sequenced, it is of interest that the deduced amino acid sequence of all of them resembles that of the vertebrate cytoplasmic actins; no gene coding for amino acid sequences typical of vertebrate muscle actin has been found⁸. Thus, the known pattern of intron sites and the available amino acid data suggest that, despite their great similarity in specialization and function, the vertebrate muscle actin and the insect muscle actin genes evolved independently from non-muscle actin genes¹⁵.

It has been suggested that split genes were assembled by rearrangement of several DNA regions containing coding sequences, into a new functional unit; noncoding flanking regions of these fragments formed the introns¹⁶. This would imply that the primordial actin gene was split in at least all those positions in which introns are found in the various actin genes, and that the variability in the location of introns was due to deletion of introns during evolution. However, without sufficient evidence, an alternative or additional mechanism should be considered also, namely that the variability in intron positions and numbers resulted from insertion and deletion of introns, possibly as transposable sequences. Studies on the structure of actin genes along the evolutionary track leading to the vertebrates may discriminate between these possibilities.

It was recently reported that a soybean actin gene contained introns at codons 20/21, 151 and 355/356 (ref. 17). When we aligned the codons of this gene with those of the rat skeletal muscle actin gene, we found that the plant actin gene intron at codon 151 and the rat skeletal muscle intron at codon 150 split the coding sequence at exactly the same position. Although we cannot exclude the possibility that the homology in intron sites in these two actin genes is coincidental, we consider this very unlikely. Thus, if these introns are evolutionarily related, the immediate conclusion is that the intron at codon 150 existed in the actin gene before the separation between the animal and plant kingdoms. Therefore, this intron cannot be considered a novel intron, specific to the skeletal muscle actin gene⁷; rather, the lack of an intron at this location in the known sea urchin actin genes and the rat β -actin gene is most probably due to a deletion. The data are compatible with the suggestion that a

primordial actin gene common to plants and animals was split at all intron locations known in the actin genes and that the variability in intron sites was a result of deletion of introns.

We acknowledge the technical assistance of Ms Zehava Levy, and thank Drs L. L. Jagodzinsky, T. Sargent and J. Bonner for the rat gene libraries. This work was supported by NIH grant R01 GM 22767 and by grants awarded by the Muscular Dystrophy Association to D.Y. and U.N. U.N. holds an A. and E. Blum Career Development Chair in Cancer Research.

Received 1 April; accepted 11 June 1982.

1. Storti, R. V. & Rich, A. *Proc. natn. Acad. Sci. U.S.A.* **73**, 2346–2350 (1976).
2. Whalen, R. G., Butler-Browne, G. S. & Gros, F. *Proc. natn. Acad. Sci. U.S.A.* **73**, 2018–2022 (1976).
3. Garrels, J. & Gibson, W. *Cell* **9**, 793–805 (1976).
4. Elzinga, M. & Lu, R. in *Contractile Systems in Nonmuscle Tissues* (eds Perry, S. V., Margreth, A. & Adelstein, R.) **29** (North-Holland, Amsterdam, 1976).
5. Vanderkerckhove, J. & Weber, K. *Differentiation* **14**, 124–133 (1979).
6. Nudel, U. *et al. Proc. natn. Acad. Sci. U.S.A.* **79**, 2763–2767 (1982).
7. Davidson, E. H. in *Genome Evolution* (eds Dover, G. A. & Flavell, R. B.) (Academic, New York, 1982).
8. Fyrberg, E. A., Bond, B. J., Hershel, N. D., Mixter, K. S. & Davidson, N. *Cell* **24**, 107–116 (1981).
9. Katcoff, D. *et al. Proc. natn. Acad. Sci. U.S.A.* **77**, 960–964 (1980).
10. Shani, M. *et al. Nucleic Acids Res.* **9**, 579–589 (1981).
11. Breathnach, R. & Chambon, P. *A. Rev. Biochem.* **50**, 349–383 (1981).
12. Weaver, R. F. & Weissmann, C. *Nucleic Acids Res.* **6**, 1175–1193 (1979).
13. Goldberg, M. thesis, Stanford Univ. (1979).
14. Ziff, E. B. & Evans, R. M. *Cell* **15**, 1463–1475 (1978).
15. Yaffe, D. *et al. in Stability and Switching in Cell Differentiation* (eds Clayton, R. N. & Truman, D. E. S.) (Plenum, New York, in the press).
16. Gilbert, W. R. *Nature* **271**, 501 (1978).
17. Shah, D. M., Hightower, R. C. & Mengher, R. B. *Proc. natn. Acad. Sci. U.S.A.* **72**, 1022–1026 (1975).
18. Firtel, R. A., Timm, R., Kimmel, A. R. & McKeown, M. *Proc. natn. Acad. Sci. U.S.A.* **76**, 6206–6210 (1979).
19. Gallwitz, D. & Sures, I. *Proc. natn. Acad. Sci. U.S.A.* **77**, 2546–2550 (1980).
20. Ng, R. & Abelson, J. *Proc. natn. Acad. Sci. U.S.A.* **77**, 3912–3916 (1980).
21. Maxam, A. M. & Gilbert, W. R. *Proc. natn. Acad. Sci. U.S.A.* **74**, 560–564 (1977).
22. Sanger, F., Nicklen, S. & Coulson, A. R. *Proc. natn. Acad. Sci. U.S.A.* **74**, 5463–5468 (1977).

Structure of the human immune interferon gene

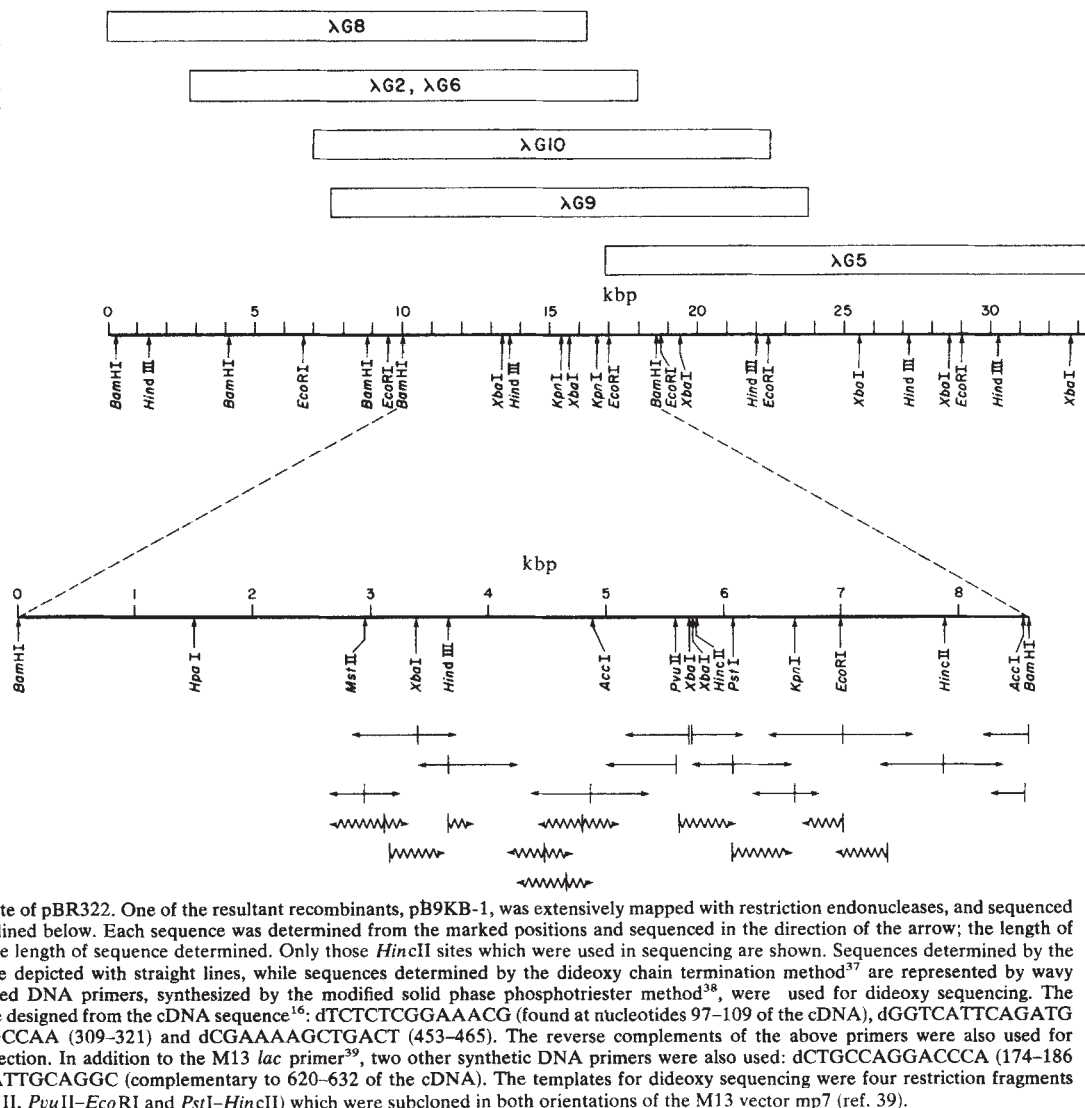
Patrick W. Gray & David V. Goeddel

Department of Molecular Biology, Genentech, Inc.,
460 Point San Bruno Boulevard, South San Francisco,
California 94080, USA

Sequence determination of cloned cDNAs and genes of the three classes of interferon (IFN- α , - β and - γ) has revealed more than a dozen members of the human IFN- α gene family^{1–7} and a single gene for IFN- β ^{4,8–14}. These genes are found on chromosome 9 (ref. 15) and contain no introns^{1,5–7,11–14}. We recently reported¹⁶ that the 146-amino acid sequence of mature IFN- γ deduced from the nucleotide sequence of a cloned cDNA was quite unrelated to those of the other IFNs, and that the gene for IFN- γ contains at least one intron¹⁶. We now describe the isolation, characterization and DNA sequence of the human IFN- γ gene. It contains three introns, a repetitive DNA element, and is not highly polymorphic. All our evidence to date^{16,17} and the present data suggest that this is the only gene for IFN- γ and that the resolution of IFN- γ into two components¹⁸ is probably the result of post-translational processing of the protein.

A ³²P-labelled restriction fragment obtained from the IFN- γ cDNA clone p69 (ref. 16) was used as a hybridization probe to screen 10⁶ plaques of a bacteriophage λ Charon 4A/human genomic library¹⁹. Six individual hybridizing plaques were isolated from the library, and DNA from each recombinant phage was isolated and mapped with five different restriction endonucleases (Fig. 1). The insert DNAs span a contiguous region of 33.5 kilobase pairs (kbp) of the human genome. Each phage DNA was analysed by Southern blotting²⁰ with the IFN- γ cDNA probe, after digestion with *Eco*RI, *Bam*HI, *Hinc*II and

Fig. 1 Restriction endonuclease map of the IFN- γ gene region. The limits of each different phage DNA insert are shown at the top. Sites for *EcoRI*, *BamHI*, *HindIII*, *KpnI* and *XbaI* were determined by digestions of each individual phage DNA. Phage (10^6) from the recombinant human λ library described by Lawn *et al.*¹⁹ were plated on to 10 15-cm plates and screened by the method of aniatas *et al.*³⁴ with 3×10^7 c.p.m. of ^{32}P -labelled³⁵ cDNA probe (887 bp *Sau3A* fragment, ref. 16). Plaques which hybridized positively with the probe were subjected to another round of plaque purification³⁴. Individual hybridizing plaques were isolated and phage DNA was prepared^{19,34} from 500-ml cultures. Each phage DNA was digested with *EcoRI*, *BamHI*, *HindIII*, *KpnI* and *XbaI* (either singly or in combination) and then electrophoresed on a 1% agarose gel. Gels were blotted on to nitrocellulose²⁰ and hybridized with the ^{32}P -labelled IFN- γ cDNA probe. The size of fragments was determined by comparison with *HindIII* λ molecular weight standards. The 8.6-kbp *BamHI* fragment of λ G9 was separated by electrophoresis through 1% agarose, electroeluted, and ligated into the *BamHI* site of pBR322. One of the resultant recombinants, pB9KB-1, was extensively mapped with restriction endonucleases, and sequenced according to the strategy outlined below. Each sequence was determined from the marked positions and sequenced in the direction of the arrow; the length of each arrow corresponds to the length of sequence determined. Only those *HincII* sites which were used in sequencing are shown. Sequences determined by the Maxam-Gilbert method³⁶ are depicted with straight lines, while sequences determined by the dideoxy chain termination method³⁷ are represented by wavy lines. Synthetic single-stranded DNA primers, synthesized by the modified solid phase phosphotriester method³⁸, were used for dideoxy sequencing. The sequences of the primers were designed from the cDNA sequence¹⁶: dTCTCTCGGAAACG (found at nucleotides 97–109 of the cDNA), dGGTCATTGAGATG (229–241), dAATGCAGAGCCAA (309–321) and dCGAAAAGCTGACT (453–465). The reverse complements of the above primers were also used for sequencing in the 3' to 5' direction. In addition to the M13 *lac* primer³⁹, two other synthetic DNA primers were also used: dCTGCCAGGACCCA (174–186 of the cDNA) and dAAATATTGAGGC (complementary to 620–632 of the cDNA). The templates for dideoxy sequencing were four restriction fragments (*HpaI*–*HindIII*, *HindIII*–*PvuII*, *PvuII*–*EcoRI* and *PstI*–*HincII*) which were subcloned in both orientations of the M13 vector mp7 (ref. 39).



PvuII. Comparison of these hybridization patterns with those obtained earlier using human lymphocyte DNA¹⁶ indicated that two phage DNAs (λ G9 and λ G10) appeared to contain the entire IFN- γ gene, as illustrated in Fig. 1. Clone λ G9 was selected for further study. The region of λ G9 hybridizing to the IFN- γ cDNA probe appeared to be contained entirely within an 8.6-kbp *BamHI* fragment. This restriction fragment was subcloned into the *BamHI* site of pBR322. The *BamHI* insert from one of the resultant recombinants, pB9KB-1, was mapped with restriction endonucleases to allow DNA sequence analysis (Fig. 1).

Almost 6,000 contiguous base pairs (bp) of the IFN- γ gene region were determined by sequencing the majority of both strands (Fig. 1); these are shown in Fig. 2. The IFN- γ cDNA sequence previously determined¹⁶ was aligned with the gene sequence as shown in Fig. 2. The two sequences matched completely except for a single base pair transition (4,712 of Fig. 2; position 588 of ref. 16) which changes the glutamine codon found at position 140 in the cDNA sequence to an arginine. However, we have subsequently sequenced five additional IFN- γ cDNAs which all contained arginine codons at this position¹⁷.

There are three intervening sequences in the coding region of the gene. All three introns begin with a GT dinucleotide and end with an AG, sequences thought to be necessary for correct RNA splicing^{21,22}. Intron 1 is 1,238 bp long; intron 2 is 95 bp, and intron 3 is 2,422 bp long. The four exons are less variable in length, coding for 38, 23, 61 and 44 amino acids, respectively.

Each intron interrupts the reading frame precisely between codons.

Although there is no obvious overall amino acid homology between IFN- γ and either IFN- α or IFN- β ¹⁶, a pattern of single amino acid homology between IFN- γ and IFN- α has been reported⁴², and the first exon of IFN- γ codes for two groups of amino acids which have homology with IFN- α . The four identical amino acids, Ser-Leu-Gly-Cys, are found at the junction between the signal peptide and mature interferons (where Cys is the first amino acid of the mature proteins)⁴. Another homology with the first exon of IFN- γ is found nearer the carboxy-terminus of IFN- α_2 (refs 2, 3):

	130	131	132	133		134	135	136
IFN- α	Tyr	Leu	Lys	Glu		Lys	Lys	Tyr
IFN- γ	Tyr	Val	Lys	Glu	Ala	Glu	Asn	Leu
	7	8	9	10	11	12	13	14

If amino acids 11–14 of IFN- γ are neglected, six of seven amino acids are identical (17 of 21-nucleotide homology). This homology may be significant as the IFN- α sequence was found to be highly conserved in this region when 11 different IFN- α sequences were compared^{4,5}. Furthermore, this region is proposed to be highly α -helical²³; the insertion of four amino acids would increase the length of the helix by one turn, but should not disturb the relative position of amino acids on the helix, which is particularly important in a postulated helix of this type

© 1982 Nature Publishing Group

Fig. 3 The IFN- γ gene region contains a repetitive element. The cloned IFN- γ gene sequence was digested with the restriction endonucleases shown and the resulting fragments (a-e) were isolated on 5% polyacrylamide gels, electroeluted and used to synthesize 32 P-labelled probes by the calf thymus primer method³⁵. The individual probes were hybridized with *Pvu*II (left) and *Eco*RI (right) digests of human lymphocyte genomic DNA (isolated by the method of Blin and Stafford⁴¹), separated on 1% agarose gels and transferred to nitrocellulose by the method of Southern²⁰. Four of the probes (a, c, d and e) hybridized only to the single *Pvu*II or *Eco*RI fragments from which they were derived. Probe b hybridized throughout the genomic DNA. The 5'- and 3'-untranslated regions of the IFN- γ gene are shown as cross-hatched boxes and coding regions of exons are represented by solid boxes.

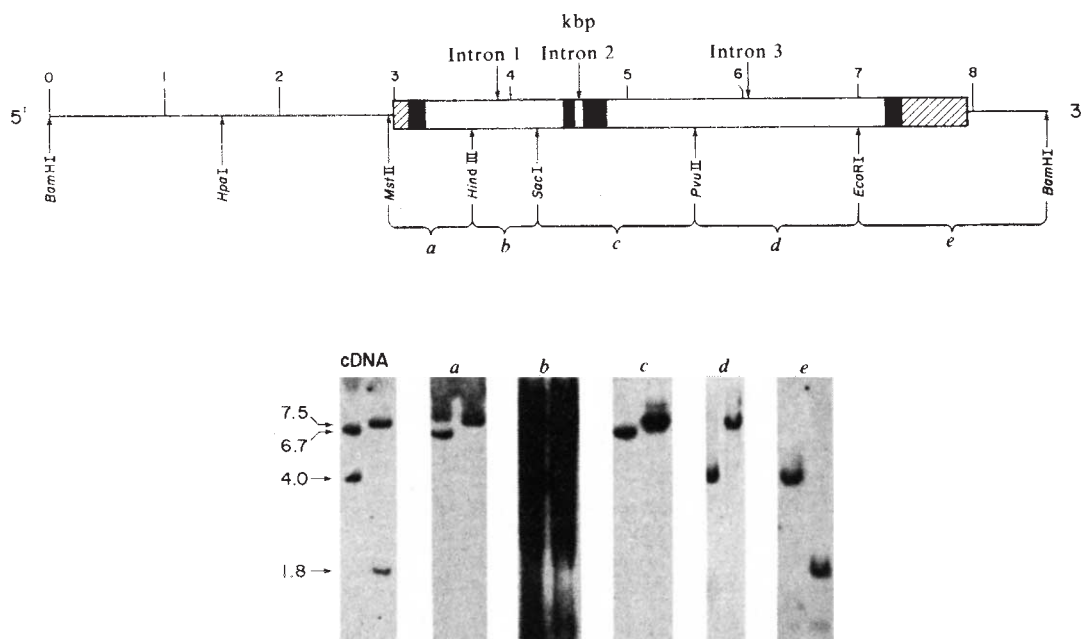
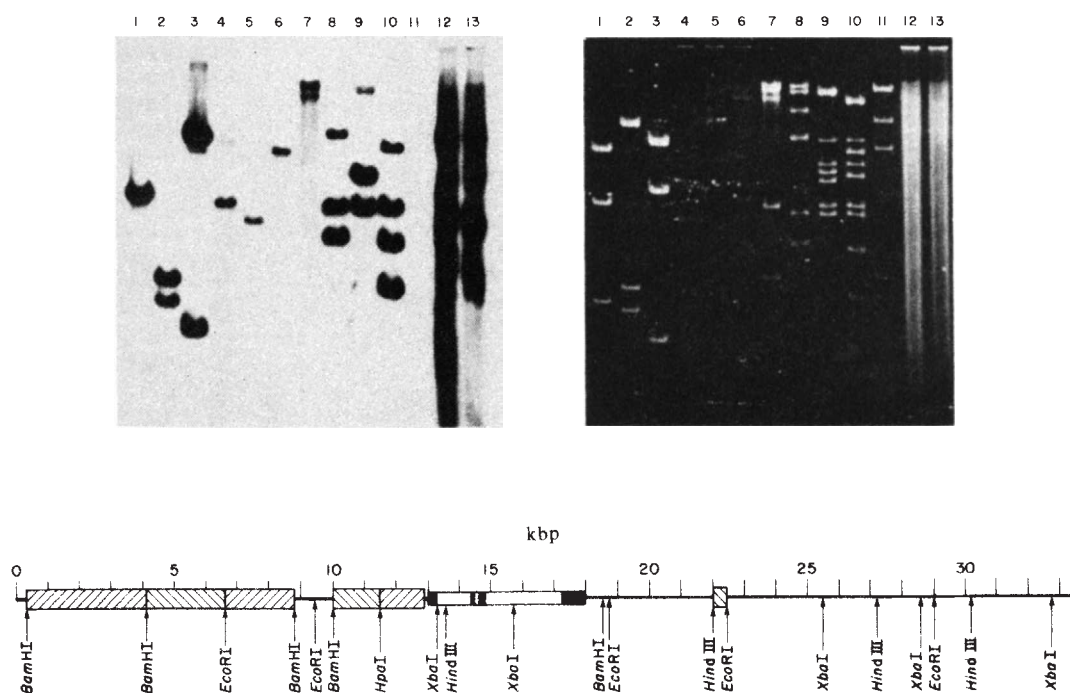


Fig. 4 Hybridization of the *Alu* family repeat probe to the IFN- γ gene region. The 32 P-labelled probe was prepared³⁵ from a 300-bp *Bam*HI fragment of pBLUR8 (provided by C. Houck and R. Lawn) which contains *Alu* family repeat DNA³². Electrophoresis in 1% agarose gel is shown on the right; lane 1, *Hind*III + *Sac*I digest of pB9KB-1; 2, *Hind*III + *Hpa*I digest of pB9KB-1; 3, *Bam*HI + *Hpa*I digest of pB9KB-1; 4, *Eco*RI digest of λ G5; 5, *Xba*I + *Hind*III digest of λ G5; 6, *Xba*I digest of λ G5; 7, *Xba*I digest of λ G9; 8, *Eco*RI digest of λ G6; 9, *Bam*HI digest of λ G8; 10, *Eco*RI + *Bam*HI digest of λ G8; 11, *Hind*III digest of λ DNA (size markers); 12, *Pvu*II digest of human genomic DNA; 13, *Eco*RI digest of human genomic DNA. The gel was transferred to nitrocellulose²⁰ and hybridized with 3×10^6 c.p.m. of probe (10^7 c.p.m. per μ g DNA), shown on the left. Four different hybridizing signals are observed in λ G8 (lane 10). The signal in the subclone pB9KB-1 was shown to consist of two separate parts (lanes 2, 3). Only a single hybridizing signal was observed in the 3' region (λ G5, lanes 4-6). Mapping of these hybridizing regions (cross-hatched) is shown below on the map originally presented in Fig. 1. The IFN- γ gene region is represented by a smaller box with solid bars for exons and open bars for introns.



We thank Mark Vasser for deoxyoligonucleotide synthesis, Dr Ellison Chen for preparation of templates for dideoxy sequencing and Dr Richard Lawn for providing the human λ DNA library.

Received 10 May; accepted 14 June 1982.

- Nagata, S., Mantei, N. & Weissmann, C. *Nature* **287**, 401-408 (1980).
- Streuli, M., Nagata, S. & Weissmann, C. *Science* **209**, 1343-1347 (1980).
- Goeddel, D. V. *et al. Nature* **287**, 411-416 (1980).
- Goeddel, D. V. *et al. Nature* **290**, 20-26 (1981).
- Lawn, R. M. *et al. Science* **212**, 1159-1162 (1981).

- Lawn, R. M. *et al. Proc. natn. Acad. Sci. U.S.A.* **78**, 5435-5439 (1981).
- Weissmann, C. *Interferon* **3**, 101-134 (1981).
- Taniguchi, T., Ohno, S., Fujii-Kuriyama, Y. & Muramatsu, M. *Gene* **10**, 11-15 (1980).
- Derynck, R. *et al. Nature* **285**, 542-547 (1980).
- Goeddel, D. V. *et al. Nucleic Acids Res.* **8**, 4057-4073 (1980).
- Lawn, R. M. *et al. Nucleic Acids Res.* **9**, 1045-1052 (1981).
- Houghton, M. *et al. Nucleic Acids Res.* **9**, 247-266 (1981).
- De Grave, W., Derynck, R., Tavernier, J., Haegeman, G. & Fiers, W. *Gene* **14**, 137-143 (1981).
- Ohno, S. & Taniguchi, T. *Proc. natn. Acad. Sci. U.S.A.* **78**, 5305-5309 (1981).
- Owerbach, D. *et al. Proc. natn. Acad. Sci. U.S.A.* **78**, 3123-3127 (1981).
- Gray, P. W. *et al. Nature* **295**, 503-508 (1982).
- Derynck, R., Leung, D., Gray, P. W. & Goeddel, D. V. *Nucleic Acids Res.* (in the press).
- Yip, Y. K., Barrowclough, B. S., Urban, C. & Vilcek, J. *Proc. natn. Acad. Sci. U.S.A.* **79**, 1820-1824 (1982).

19. Lawn, R. M., Fritsch, E. F., Parker, R. C., Blake, G. & Maniatis, T. *Cell* **15**, 1157-1174 (1978).
20. Southern, E. M. *J. molec. Biol.* **98**, 503-517 (1975).
21. Breathnach, R., Benoist, C., O'Hare, K., Gannon, F. & Chambon, P. *Proc. natn. Acad. Sci. U.S.A.* **75**, 4853-4857 (1978).
22. Lerner, M. R. *et al. Nature* **283**, 220-224 (1980).
23. Wetzel, R. *et al. in Chemistry and Biology of Interferons: Relationship to Therapeutics* (eds Merigan, T. C., Friedman, R. M. & Fox, C. F.) (Academic, New York, in the press).
24. Goldberg, M. thesis, Stanford Univ. (1979).
25. Gannon, F. *et al. Nature* **278**, 428-434 (1979).
26. Grosschedl, R. & Birnstiel, M. L. *Proc. natn. Acad. Sci. U.S.A.* **77**, 1432-1436 (1980).
27. Waslyk, B. *et al. Proc. natn. Acad. Sci. U.S.A.* **77**, 7024-7028 (1980).
28. Grosveld, G. C., Shewmaker, C. K., Jat, P. & Flavell, R. A. *Cell* **25**, 215-226 (1981).
29. Benoist, C., O'Hare, K., Breathnach, R. & Chambon, P. *Nucleic Acids Res.* **8**, 127-142 (1980).
30. Grosveld, G. C., de Boer, E., Shewmaker, C. K. & Flavell, R. A. *Nature* **295**, 120-126 (1982).
31. Houck, C. M., Rinehart, F. P. & Schmid, C. W. *J. molec. Biol.* **132**, 289-306 (1979).
32. Rubin, C. M., Houck, C. M., Deininger, P. L., Friedman, T. & Schmid, C. W. *Nature* **284**, 372-374 (1980).
33. Calabretta, B. *et al. Nature* **296**, 219-225 (1982).
34. Maniatis, T. *et al. Cell* **15**, 687-701 (1978).
35. Taylor, J. M., Illmensee, R. & Summers, S. *Biochim. biophys. Acta* **442**, 324-330 (1976).
36. Maxam, A. M. & Gilbert, W. *Meth. Enzym.* **65**, 499-560 (1980).
37. Smith, A. J. H. *Meth. Enzym.* **65**, 560-580 (1980).
38. Crea, R. & Horn, T. *Nucleic Acids Res.* **8**, 2331-2348 (1980).
39. Gressing, J., Crea, R. & Seeburg, P. H. *Nucleic Acids Res.* **9**, 309-321 (1981).
40. Proudfoot, N. J. & Brownlee, G. G. *Nature* **263**, 211-214 (1976).
41. Blin, N. & Stafford, D. W. *Nucleic Acids Res.* **3**, 2303-2308 (1976).
42. Epstein, L. B. *Nature* **295**, 453-454 (1982).

DNA precursors in chemical mutagenesis: a novel application of DNA sequencing

Michael D. Topal*, Clyde A. Hutchison III† & Mary S. Baker*

Departments of *Pathology, †Bacteriology and Immunology, and the *Cancer Research Centre, University of North Carolina Medical School, Chapel Hill, North Carolina 27514, USA

Recently, we have shown that deoxyribonucleoside residues in the cellular DNA precursor pool are generally more susceptible to methylation than are residues within the DNA duplex¹. The N-1 position of adenosine, for example, was found to be at least 13,000 times more susceptible to methylation by *N*-methyl-*N*-nitrosourea (MNU) than the same site in the DNA. These results suggest that potential sites for alkylation in the double-strand duplex are relatively inaccessible to direct alkylation *in vivo*. Many of these sites are probably protected from alkylation not only by their position in the interstices of the DNA helix^{2,3}, but also by further *in vivo* 'packaging' of the DNA in chromatin⁴⁻⁸. We have now used DNA sequencing to demonstrate the incorporation properties of products of the reaction of MNU with dATP and of deoxy-*N*⁴-hydroxycytidine triphosphate during DNA replication *in vitro* by phage T4 DNA polymerase and the 'Klenow' fragment of *Escherichia coli* pol I. The results suggest that DNA precursor nucleotides due to their greater availability for alkylation, may offer routes for the introduction of alkylated residues into double-stranded DNA.

To investigate the ability of DNA polymerase to use products from the interaction of chemical mutagens with DNA precursors, we reacted MNU with dATP. The products were separated by chromatography on QAE-Sephadex. Chromatographic and enzymatic analyses (to be described elsewhere) revealed N-1 and N-3MedATP in an approximate ratio of 7:1, N-1/N-3MedATP, dADP with one methyl group associated with its terminal β -phosphate, mP ^{β} -dADP, and dATP with one methyl group associated with its terminal γ -phosphate, mP ^{γ} -dATP.

We investigated the incorporation of these products into DNA using a novel application of the plus and minus' DNA sequencing method⁹. A plus reaction with a modified dNTP, dN*TP, as substrate provides information about the incorporation properties of dN*TP when bands generated on an analysing gel by products of the reaction are compared with those generated on the same gel by products of control reactions

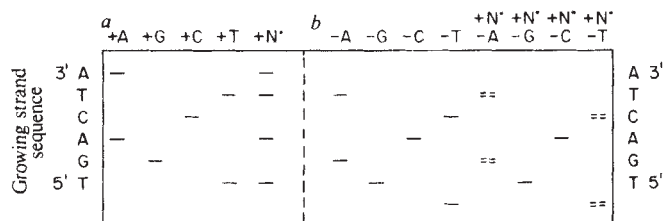


Fig. 1 Representation of a gel containing the products of: *a*, plus reactions; and *b*, minus reactions for the DNA sequence TGACTA. The gel bands indicated are the products of plus and minus reactions in the presence of a hypothetical modified triphosphate, dN*TP (N*), which incorporates as dATP and TTP (that is, against T and A template residues). Note the appearance of bands in the +N* slot that correspond to bands in the +A and +T but not the +G or +C slots. In minus reactions, bands in slots that analyse the products of -A and -T reactions with N* have decreased in intensity or are absent (indicated by pairs of dotted lines) compared with bands from -A and -T reactions without N*. No such effect was seen when the products of -G and -C reactions in the presence and absence of N* were compared.

from the same sequence (Fig. 1*a*). The incorporation properties of dN*TP can be determined also from minus reactions using dN*TP instead of the missing triphosphate. If dN*TP can substitute for the missing dNTP, then synthesis proceeds beyond those sites that are blocked in the control reaction because of the need for the missing dNTP. Incorporation of the methylated precursor eliminates bands in the gel (see Fig. 1*b*).

*N*⁴-hydroxy-dCTP (OH-dCTP; IUPAC numbering) was studied as a control in both plus and minus reactions (Fig. 2) because it may be able to act as both dCTP and TTP as suggested by studies of the incorporation of OH-rCTP into RNA^{10,11}. As expected, OH-dCTP incorporated sometimes as dCTP and sometimes as TTP but never as dATP or dGTP. In a plus reaction, however, OH-dCTP incorporated opposite less than half of the available A and G sites (Fig. 2*a*) whereas in minus reactions containing OH-dCTP, almost all A and G template sites were used (Fig. 2*b*). This apparently different ability to incorporate in plus as opposed to minus reactions is probably related to the heavy dependence of plus reactions on the exonuclease and synthetic activities of the DNA polymerase. Once a residue (correct or incorrect) is incorporated during a minus reaction it may escape repeated scrutiny by the proofreading exonuclease through incorporation of the next residue. Analysis of OH-dCTP incorporation in plus reactions (Table 1) suggests

Table 1 Incorporation of OH-dCTP as T and C during plus reactions

<i>a</i> Four sites incorporate OH-dCTP at:			Two do not incorporate OH-dCTP at:		
Amount of incorporation	Site	ΦX174 coordinate	Site	ΦX174 coordinate	
(m)	CT:C:GG	(4,683)	AG:C:AA	(4,677)	
(s)	CC:C:TC	(4,685)	GG:C:GT	(4,710)	
(w)	GC:C:TT	(4,694)			
(w)	TA:C:TG ₃	(4,724)			
<i>b</i> Two sites incorporate OH-dCTP at:			Five do not incorporate OH-dCTP at:		
Amount of incorporation	Site	ΦX174 coordinate	Site	ΦX174 coordinate	
(w)	CC:T:CG	(4,684)	CT:T:GC	(4,692)	
(m)	AC:T:GA ₃	(4,723)	AT:T:AG	(4,698)	
			TG:T:GA	(4,706)	
			CG:T:GT	(4,708)	
			GA:T:AG ₃	(4,720)	

a, Six C sites and *b*, seven T sites tested for incorporation. The site of incorporation is indicated by the ΦX174 *am3cs70* plus strand DNA sequence coordinate¹⁸ shown next to each sequence. The minus (daughter) strand site of incorporation is indicated by the ΦX174 *am3cs70* plus strand DNA sequence neighbours and next-nearest minus strand neighbours are also shown. A semi-quantitative measure of the ability of OH-dCTP to incorporate at the site indicated is represented by w, m or s (weak, moderate or strong), a measure of the effect on band intensity.

that the extent to which OH-dCTP can replace dCTP or TTP during DNA synthesis *in vitro* varies with sequence position; OH-dCTP apparently preferentially incorporates as dCTP when at least one of its nearest neighbours is T and as TTP when at least one of its nearest neighbours is C. The presence of purine residues on both sides of C or T sites inhibits OH-dCTP incorporation opposite those G or A template sites.

Analysis of the products of reaction of MNU with dATP by the plus reaction method (Fig. 3a) indicates that N-1/N-

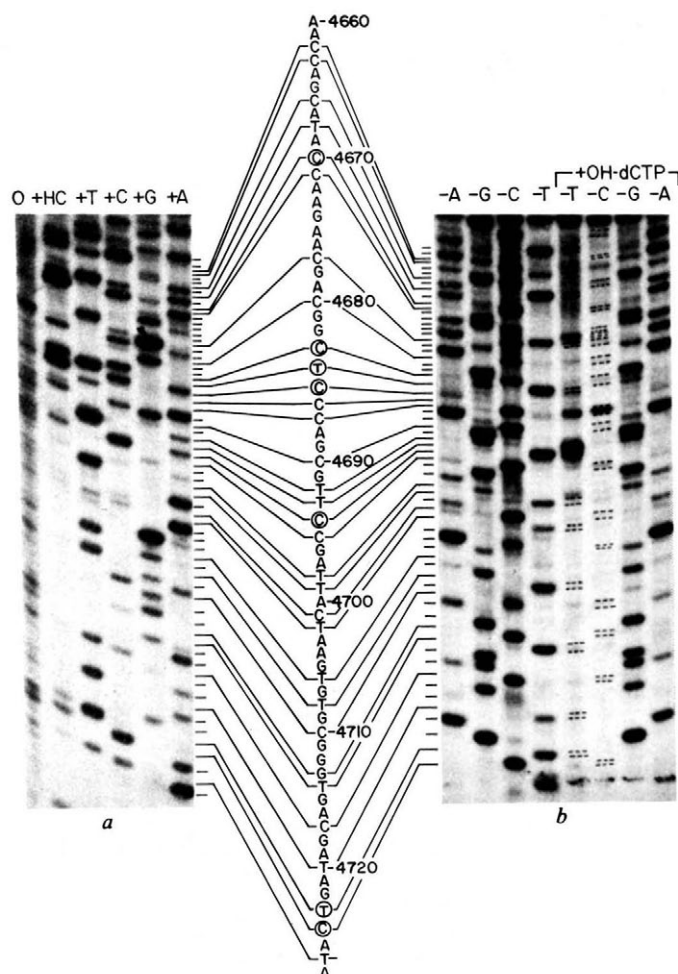


Fig. 2 Incorporation of OH-dCTP (HC) during: *a*, plus and *b*, minus sequencing reactions. OH-dCTP was synthesized according to published procedures¹⁹, shown to be pure by HPLC, and identified by UV absorbance. Plus strand Φ X174 *am3 cs70* DNA annealed with the *Hae*III restriction fragment Z9 (prepared as described elsewhere²⁰) was incubated at 37°C for 20 min with 0.4 μ g of bacteriophage T4 DNA polymerase in the presence of 0.2 mM dNTP for plus reactions, or 10.0 μ M dNTP for minus reactions. OH-dCTP (500 μ M) was used where indicated. *a*, Products of normal plus reactions and products of a plus reaction in the presence of OH-dCTP are analysed on this gel. The sequence of the minus strand (growing strand) is shown. Sites at which OH-dCTP incorporated during plus reactions are indicated by circled bases in the sequence. Only those sites where incorporation was unambiguously assigned are indicated. Caution is required analysing this gel because the bands slant downwards as you proceed from left to right across the gel. *b*, Products of normal minus reactions and products of four extra minus reactions in the presence of OH-dCTP. Bands that have been diminished in intensity by the presence of OH-dCTP during minus reactions are indicated by pairs of dotted lines. The control slot, labelled O, contains the protein-free products of the asynchronous extension of template-primer (by Klenow pol I) in the absence of phage T4 DNA polymerase. All sequence sites are represented by radio-labelled primer extensions. These extended primers are either hydrolysed or elongated by DNA polymerase during plus and minus reactions, respectively.

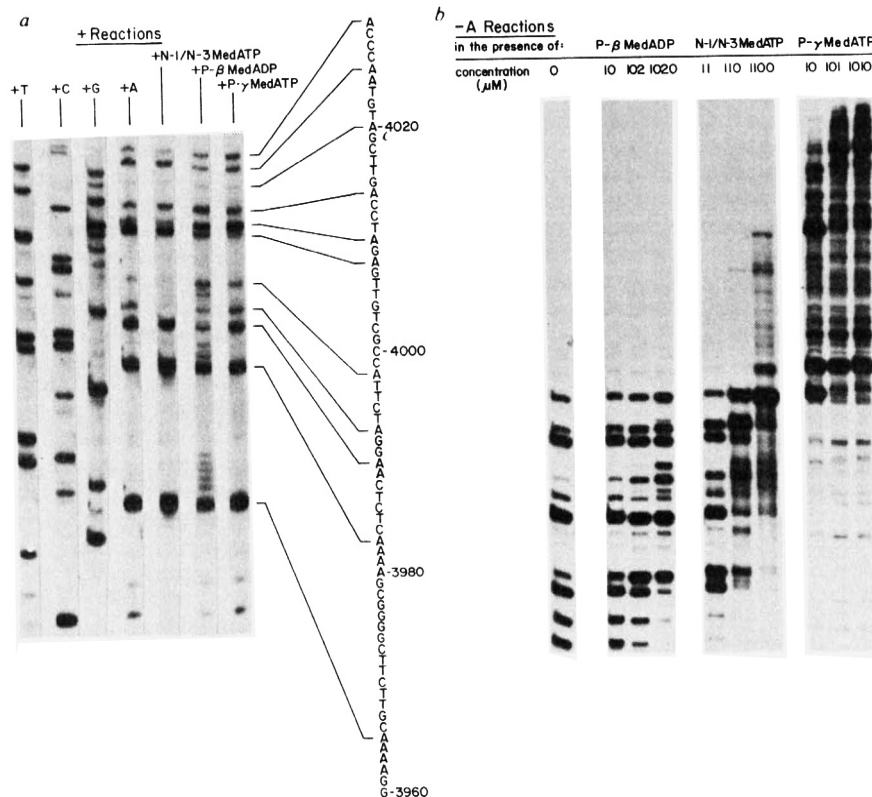
3MedATP, mP^γ-dATP and mP^β-dADP are incorporated into DNA by bacteriophage T4 DNA polymerase opposite T template residues but not opposite A, C or G template residues (not shown) at the concentration used. Incorporation of mP^β-dADP opposite T is suspect, however, for two reasons. First, it is believed that diphosphates cannot be incorporated by DNA polymerase. Second, this product chromatographs very closely to unreacted dATP, and so is possibly contaminated by dATP. All of the products were therefore re-purified until the diphosphate analogue no longer incorporated into DNA. Repurification did not alter the ability of products other than mP^β-dADP to pair as dATP. Use of repurified methylation products in minus reactions confirms that N-1/N-3MedATP and mP^γ-dATP preferentially incorporate opposite T template residues. Generally, mP^β-dADP at 10, 102 and 1,020 μ M did not incorporate opposite any template residues (Fig. 3b); this result serves as a control against contamination by dATP. The slight incorporation at 1,020 μ M suggests very low levels (< 4 nM) of dATP contamination. Yet lower levels of dATP contamination are expected for the other products as they chromatographed even further away from the unreacted dATP.

N-1/N-3MedATP incorporates into DNA *in vitro* at T template sites (Fig. 3b) but not at A, G or C sites (not shown) at the concentrations used. Incorporation at some T template sites was noted at 10 μ M and efficient incorporation was attained at 1,010 μ M as demonstrated by the conversion of lower to higher molecular weight DNA fragments. In contrast, dATP incorporates at some sites at 4 nM and incorporates efficiently at 500 nM. This suggests that one N-1MedATP will incorporate for every 100–1,000 dATPs if both are present at equal concentrations during DNA replication. The use of chemically synthesized N-1MedATP in minus A reactions (unpublished results) indicates that it completely mimics the action of the N-1/N-3MedATP mix in a minus A reaction. Whether or not N-3MedATP can incorporate as dATP therefore remains unknown. mP^γ-dATP incorporates well opposite T (Fig. 3b) but not opposite A, G or C residues (not shown). Incorporation of mP^γ-dATP caused a shift of almost all the lower molecular weight DNA fragments, formed because of the absence of dATP during DNA synthesis, to higher molecular weight even at 10 μ M—behaviour similar to that found with 10 μ M dATP. Thus, the presence of a methyl adduct on the terminal phosphate of dATP does not interfere significantly with incorporation of this precursor.

In a preliminary report we suggested that some unpurified and unidentified products of the reaction of MNU with dATP incorporate as dGTP and dCTP as well as dATP at high concentrations¹². The data presented here show that at much greater purity, those products that we have been able to isolate and clearly identify incorporate preferentially opposite T template residues.

Alkylation of DNA precursors may result in nucleotide pool imbalances, inhibition of nucleotide binding enzymes, and incorporation of modified residues into double-stranded DNA *in vivo*. The plus and minus DNA sequencing technique has proved a valuable tool in elucidating the third possibility. The technique requires only microgramme quantities of modified dNTP, does not require radiolabelling of the modified dNTP of interest, and allows us to locate specific sites where modified dNTPs are incorporated into DNA. By using this technique we have shown that N-1MedATP incorporates as dATP during replication of single-stranded DNA (ssDNA) *in vitro* by phage T4 DNA polymerase. Preliminary studies in our laboratory further indicate that the Klenow fragment of *Escherichia coli* pol I also incorporates N-1MedATP opposite T template residues. Similar *in vivo* conversion of N-1MedATP to N-1MedAMP template residues could be mutagenic as such template residues have been shown to cause misincorporation of C, G, U and A residues during 'transcription' of RNA with RNA polymerase¹³. Indeed, alkylation of unpaired DNA residues in the form of precursors and possibly single-strand

Fig. 3 Incorporation properties of products from the reaction of MNU with dATP during: *a*, plus and *b*, minus reactions. The template-primer is a synthetic oligodeoxyribonucleotide (P-L Biochemicals) complementary to positions 3,466 to 3,477 (pCAATAAACAGC) of the pBR322 insert of R208, a pBR322/f1 chimera²¹, annealed to R208 ssDNA. The replicating daughter-strand sequence shown agrees with the published sequence for pBR322 (ref. 22). Reaction conditions are as described in Fig. 2 legend. *a*, Plus reactions: the concentrations of modified precursors were all 1.0 mM. All three products incorporated as dATP, as shown by the appearance of bands corresponding to A sites in the sequence (denoted by lines connecting bands with sequence sites). On further purification of the products, mP^β-dADP(P-βMedAP) lost its ability to incorporate. The other repurified products continued to incorporate as dATP, however, making the diphosphate analogue a good control against dATP contamination. These further purified products were used in minus reactions.



b, The template-primer and reaction conditions were the same as for *a*. Note the increased shifting of bands to higher positions in the gels as you proceed from the control gel slot on the left to the highest concentration of mP^γ-dATP(P-γMedATP) at the extreme right. This shifting of bands indicates incorporation into DNA opposite T template residues.

residues at the replication fork, may offer a direct explanation for the S-phase dependence of both mutation¹⁴ and neoplastic transformation¹⁵ of cells in culture and of the chromosome replication point specificity of alkylating agents in bacteria^{16,17}.

This work was supported by USPHS grants to M.D.T. (GM 24798 and CA 28632), C.A.H. (GM 21313 and AI 08998) and M.S.B. (predoctoral training grant ES 07017).

Received 5 May; accepted 19 June 1982.

Antifreeze effect of thermal hysteresis agents protects highly supercooled insects

Karl Erik Zachariassen & Jon Arve Husby

Department of Zoology, University of Trondheim, 7055 Dragvoll, Norway

Insects that are cold-hardy but sensitive to freezing survive at low temperatures due to their capacity for supercooling—their supercooling points are below -25°C . In addition to polyols, these insects produce proteinaceous thermal hysteresis antifreeze agents reported to prevent growth of ice crystals down to $\sim -10^{\circ}\text{C}$, thus stabilizing the supercooled insects down to this temperature^{1,2}. We report here that in the beetle *Rhagium inquisitor* the ability of thermal hysteresis agents to prevent growth of ice crystals increases substantially when the crystal size is diminished. Thus, supercooled insects may be protected against growth of embryo ice crystals over their whole supercooling range, that is, they may be protected against freezing even during prolonged exposures to temperatures as low as -30°C . The present observations are in accord with the prevalent theories for the action of antifreeze agents.

Thermal hysteresis factors (THF) have been found in polar fish^{3,4}, molluscs⁵ and insects^{1,2}. They appear to act by causing a separation of melting point and freezing point (thermal hysteresis) of frozen samples of body fluid, so that ice crystals are prevented from growing on cooling until a temperature several degrees below the melting point is reached. At this temperature (the freezing point) a rapid growth of ice crystals is observed. Thermal hysteresis factors are present in cold-hardy freeze-

- Topal, M. D. & Baker, M. S. *Proc. natn. Acad. Sci. U.S.A.* **79**, 2211–2215 (1982).
- Shapiro, R., Braverman, B., Louis, J. B. & Servis, R. B. *J. biol. Chem.* **248**, 4060–4064 (1973).
- Cramer, F. *Prog. Nucleic Acid Res. molec. Biol.* **11**, 391–421 (1973).
- Kolchinsky, A. M., Mirzabekov, A. D., Gilbert, W. & Li, L. *Nucleic Acids Res.* **3**, 11–18 (1976).
- Rajalakshmi, S., Rao, P. M. & Sarma, D. S. R. *Biochemistry* **17**, 4515–4518 (1978).
- Moses, H. L., Webster, R. A., Martin, G. D. & Spelsberg, T. C. *Cancer Res.* **36**, 2905–2910 (1976).
- Ramanathan, R., Rajalakshmi, S., Sarma, D. S. R. & Farber, E. *Cancer Res.* **36**, 2073–2079 (1976).
- Bailey, G. S., Nixon, J. E., Hendricks, J. D., Sinnhuber, R. O. & Van Holde, K. E. *Biochemistry* **19**, 5836–5842 (1980).
- Sanger, F. & Coulson, A. R. *J. molec. Biol.* **94**, 441–448 (1975).
- Kochetkov, N. K. & Budowsky, E. I. *Prog. Nucleic. Acid. Res. molec. Biol.* **9**, 403–438 (1969).
- Muller, W., Weber, H., Meyer, F. & Weissman, C. *J. molec. Biol.* **124**, 343–358 (1978).
- Topal, M. D., Hutchinson, C. A. III, Baker, M. S. & Harris, C. *ICN-UCLA Symp. molec. cell. Biol.* **19**, 725–734 (1980).
- Kroger, M. & Singer, B. *Biochemistry* **18**, 3493–3500 (1979).
- Tong, C., Fazio, M. & Williams, G. M. *Proc. natn. Acad. Sci. U.S.A.* **77**, 7377–7379 (1980).
- Grisham, J. W., Greenberg, D. S., Kaufman, D. G. & Smith, G. *Proc. natn. Acad. Sci. U.S.A.* **77**, 4813–4817 (1980).
- Cerda-Olmeda, E., Hanawalt, P. C. & Guerola, N. *J. molec. Biol.* **33**, 705–719 (1968).
- Hince, T. A. & Neale, S. *Mutat. Res.* **24**, 383–387 (1974).
- Sanger, F. *et al.* *J. molec. Biol.* **125**, 225–246 (1978).
- Budowsky, E. J., Shibaeva, R. P., Sverdlov, E. D. & Monastyrskaya, G. S. *Molec. Biol. USSR* **2**, 321–328 (1968).
- Lautenberger, J. A. *et al.* *Proc. natn. Acad. Sci. U.S.A.* **75**, 2271–2275 (1978).
- Boeke, J. D., Vovis, G. F. & Zinder, N. D. *Proc. natn. Acad. Sci. U.S.A.* **76**, 2699–2702 (1979).
- Sutcliffe, J. G. *Cold Spring Harb. Symp. quant. Biol.* **43**, 77–90 (1979).

sensitive beetles of many species, giving their body fluid freezing points down to $\sim -10^\circ\text{C}$. However, all these beetles have supercooling points below -20°C , suggesting that THF are of no protective value to the insects in the critical, highly supercooled state.

Preliminary experiments revealed that diminution of ice crystals was accompanied by a more pronounced thermal hysteresis effect. To investigate the exact relationship between ice crystal size and thermal hysteresis, experiments were done on haemolymph from cold-hardy freeze-sensitive cerambycid beetles, *R. inquisitor*. These beetles hibernate at the adult stage

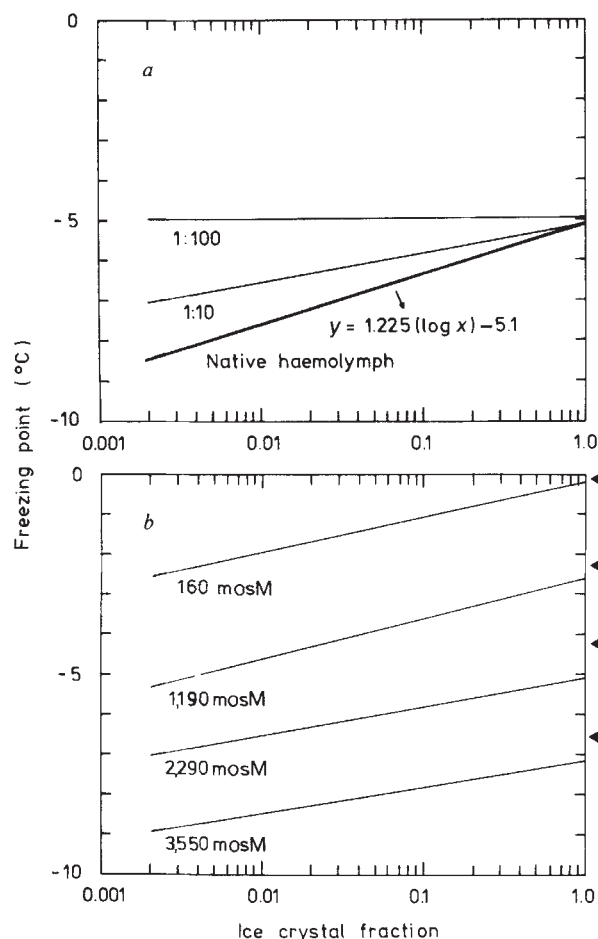


Fig. 1 a, Freezing point of native haemolymph of *R. inquisitor* beetles (thick line) and of haemolymph diluted 10-fold and 100-fold in isosmotic (2,290 mosM) glycerol solution (thin lines) as a function of ice crystal size. b, Freezing points of haemolymph of *R. inquisitor* beetles diluted 10-fold in four different concentrations of glycerol solution as a function of ice crystal size. The crystal size is expressed as the ice crystal fraction of the whole (2 nl) fluid sample. Assuming that the osmotic coefficients remain constant during freezing, the fraction of unfrozen water in a partly frozen sample is given by t_m/t_i , where t_m is the melting point of the fluid sample and t_i the temperature at which the ice crystal is stabilized. Thus, the ice crystal fraction has been calculated as $1 - t_m/t_i$. The linear regression lines were each calculated from ~ 75 individual points. The correlation coefficients range from 0.53 to 0.77, and the s.d. about the regression lines is about $\pm 0.6^\circ\text{C}$; the comparatively large s.d. may, at least in part, be due to the fact that the ice crystals were not ideal spheres. The experiments were carried out with a Clifton nanolitre osmometer, where 2 nl fluid samples could be observed through a microscope, while the temperature was regulated to within $\pm 0.001^\circ\text{C}$ between 0°C and -9.5°C . The beetles were collected in the vicinity of Trondheim in January 1981. The dilutions were carried out with the haemolymph and glycerol solutions sandwiched between two layers of paraffin oil inside glass capillaries. Fluid was transferred between the capillaries and to the osmometer by means of a microsyringe.

and incorporate accumulated glycerol and THF, thus depressing body fluid melting points to below -5°C and supercooling points of intact animals to below -25°C . Our experiments revealed (Fig. 1a) that there is a linear relationship between the freezing points and the logarithm of the ice crystal size over the whole range of the measurements, the freezing point corresponding to the smallest ice crystal tested being about -9°C .

Even the smallest ice crystal used in these experiments was large compared with the aggregates of water molecules present in the body fluid of a highly supercooled insect. Figure 1a indicates that the freezing points will continue to decrease with further diminution of the ice crystal. Hence, THF may be able to prevent ice crystal growth and thus stabilize supercooled insects at considerably lower temperatures than those tested experimentally. Unfortunately, the available techniques for temperature control do not allow stabilization of ice crystals at a smaller size. However, some information can be obtained by extrapolating the regression line representing pure haemolymph in Fig. 1 to smaller crystal sizes. The haemolymph used in the experiment was 2,290 mosM. *R. inquisitor* beetles with this haemolymph osmolality have supercooling points of $\sim -25^\circ\text{C}$ (ref. 6). Extrapolation of the regression line to this temperature reveals that the corresponding ice crystal fraction is $\sim 10^{-16}$, that is, an ice crystal radius of 3.6×10^{-17} mm (fluid sample volume = 2 nl). According to Chalmers⁷, the critical crystal radius for nucleation by embryo ice crystals at -25°C is 1.5×10^{-16} mm. The aggregations of water molecules in a supercooled insect are probably smaller than this critical size. The extrapolated regression line indicates that THF have the ability to prevent growth of embryo ice crystals of this size. Thus, THF seem to have considerably more efficient antifreeze effects at low temperatures than previously thought. They may in fact be able to stabilize supercooled insects over their whole supercooling range. Thus, the presence of THF probably explains why many hibernating insects can endure exposure to extreme cold for long periods without freezing.

The influence of varying concentrations of THF on the thermal hysteresis was investigated in another series of experiments, in which the concentration of THF was varied by diluting *R. inquisitor* haemolymph 10-fold and 100-fold in isosmotic (2,290 mosM) glycerol solution. These studies revealed (Fig. 1a) that even after 10-fold dilution there was still a pronounced thermal hysteresis effect, but this was absent following 100-fold dilution. Seasonal variations in the levels of THF may influence the freezing points in accordance with the varying slopes of the regression lines in Fig. 1a.

The fact that a pronounced thermal hysteresis was still present after 10-fold dilution of the haemolymph was used to investigate the relationship between glycerol concentration and thermal hysteresis (Fig. 1b). The results reveal that increasing glycerol concentrations cause a parallel displacement of the regression lines to lower freezing points, the freezing point depressions being greater than the corresponding melting point depressions. Thus, glycerol seems to act by extending the temperature range within which the THF performs its stabilizing antifreeze effect.

As THF appear to be unable to mask the effect of ice nucleating agents⁸, their antifreeze function at very low temperatures is restricted to insects which have no nucleating agents in their body fluids^{6,9,10}.

The present results are in agreement with the model proposed by Raymond and DeVries^{11,12} to describe the mechanism of the THF antifreeze effect. This model is based on the assumption that THF molecules are adsorbed to the ice crystal surface, thus confining ice growth to the spaces between adjacent THF molecules. The resulting increase in ice front curvature will displace ice growth to lower temperatures. When the ice crystal size is diminished as in the present experiments, more THF molecules may be adsorbed per unit ice surface area. The subsequent increase in ice front curvature will therefore cause a further decrease in the freezing points, as observed.

Received 8 April; accepted 1 July 1982.

1. Duman, J. G. *J. Insect Physiol.* **25**, 805–810 (1979).
2. Husby, J. A. & Zachariassen, K. E. *Experientia* **36**, 963–964 (1980).
3. DeVries, A. L. *Science* **172**, 1152–1155 (1971).
4. DeVries, A. L., Komatsu, S. K. & Feeney, R. E. *J. biol. Chem.* **245**, 2901–2908 (1970).
5. Theede, H., Schneppenheim, R. & Beress, L. *Mar. Biol.* **36**, 183–186 (1976).
6. Zachariassen, K. E. *J. comp. Physiol.* **140**, 227–234 (1980).
7. Chalmers, B. *Scient. Am.* **200** (2), 114–122 (1959).
8. Baust, J. G. & Zachariassen, K. E. *Cryo. Lett.* (submitted).
9. Zachariassen, K. E. & Hammel, H. T. *Nature* **262**, 285–287 (1976).
10. Sømme, L. & Conradi-Larsen, E. M. *Oikos* **29**, 118–126 (1977).
11. Raymond, J. A. & DeVries, A. L. *Proc. natn. Acad. Sci. U.S.A.* **74**, 2589–2693 (1977).
12. DeVries, A. L. in *Animals and Environmental Fitness* (ed. Gilles, R.) 583–607 (Pergamon, Oxford and New York, 1980).

E. coli F₁-ATPase interacts with a membrane protein component of a proton channel

John E. Walker, Matti Saraste & Nicholas J. Gay

Laboratory of Molecular Biology, MRC Centre, Hills Road, Cambridge CB2 2QH, UK

The ATP synthases of bacteria, mitochondria and chloroplasts, which use the energy of a transmembrane proton gradient to power the synthesis of ATP, consist of an integral membrane component F₀—thought to contain a proton channel—and a catalytic component, F₁. To help investigate the way F₀ and F₁ are coupled, we have sequenced the b-subunit of the *Escherichia coli* F₀, which seems to be the counterpart of a thermophilic bacteria F₀ subunit thought to be essential for F₁ binding¹. We report here that its sequence is remarkable, being hydrophobic around the N-terminus and highly charged in the remainder. We propose that the N-terminal segment lies in the membrane and the rest outside. The extramembranous section contains two adjacent stretches of 31 amino acids where the sequence is very similar: in the second of these stretches there is further internal homology. These duplicated stretches of the polypeptide probably fold into two α -helices which have many common features able to make contact with F₁ subunits. Thus protein b occupies a central position in the enzyme, where it may be involved in proton translocation. It is possibly also important in biosynthetic assembly.

The organization of the genes coding for ATP-synthase subunits in the *unc* operon is shown in Fig. 1. The three genes for F₀ polypeptides a, c and b are grouped together at the promoter-proximal end and are followed by the five F₁ genes. Gene b encodes a protein of molecular weight 17,212, which is not modified by post-translational processing². The amino acid sequence of this product (Fig. 2) falls into two distinct parts. With the exception of single threonine and cysteine residues and one charged residue, Lys 23, residues 1–33 are hydrophobic. It is proposed that this section of the protein forms a single transmembrane α -helix (see below). Other transmembrane α -helices in bacteriorhodopsin also contain buried threonine residues³. The hydroxyl group can form a hydrogen bond with the main-chain polypeptide backbone, as can the thiol of cysteine. Buried charges are also proposed in bacteriorhodopsin but in protein b the side chain of Lys 23 could be near the membrane surface. In contrast, the rest of the sequence of protein b is highly polar, containing 26 basic and 26 acidic residues; this charged section of the protein has internal repeats (Fig. 3). Thus, residues 52–82 are homologous with residues 85–105 (Fig. 4). In addition, within the latter but not the former sequence, a shorter repeat is found, residues 85–99 being homologous to residues 100–115. It is of particular interest that the residues identical in the short repeat consist largely of those that are not conserved (10/14) in the long repeat (see Fig. 4). Thus it seems that part of the charged segment arose by duplication of a 31-amino acid stretch which itself was formed in an earlier event involving duplication of 15 amino acids.

Further information about the structure of protein b can be obtained from the secondary structure predicted from the amino

acid sequence (Fig. 1). It is highly probable (probability >0.6) that the protein is almost entirely α -helical. Breaks in α -helices may occur at three points, around residues 27–28 (Pro-Pro), 83–86 (between the long repeats) and near residue 136. A weakening of α -helicity is also noted at residues 43–44. Thus the structure of the protein breaks down into two clear α -helical domains, the first of which (residues 1–20) is highly hydrophobic and could yield an α -helix spanning the membrane once. This helix presumably makes hydrophobic interactions with a and c, the two other constituents of F₀. It is hinged to the second highly charged region by residues 21–28, including the sequence Pro-Pro. The highly charged region appears to consist of at least two connected α -helices, with related sequences, and presumably presenting α -helical surfaces with many identical features able to interact with F₁ subunits. A clue to the nature of the interactions between the proposed α -helical structure of b and F₁ comes from a study of observed helix-helix interactions in proteins⁴, which showed that Ala, Val, Leu, Ile and Phe constitute 50% of all residues involved in these interactions, and Asn, Gln, Glu, Asp, Lys and Arg a further 25%, in contrast to the residues involved in helix- α -sheet interactions⁵. The proposed extramembranous α -helical region of b consists mainly of Ala, Val, Leu and Ile (44%) and acidic and basic amino acids (39%). Thus, if the rules of protein-protein interactions are the same as found in proteins to date, it seems likely

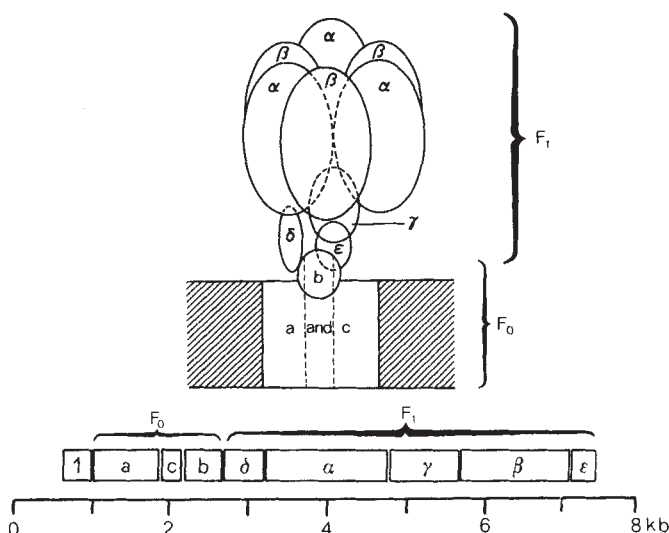


Fig. 1 Organization of polypeptides in the ATP-synthase complex and of their corresponding genes in the *unc* operon. The enzyme (upper panel) is bound to the inner membrane of *E. coli*. F₁ protrudes into the cytoplasm and is depicted with alternating α - and β -subunits although other arrangements have been proposed^{13,18}. The hydrophobic F₀ subunits, a and c, are buried in the membrane. It is proposed here that a and c interact with the hydrophobic amino-terminal region of the third F₀ subunit, b. The rest of b is very polar and lies outside the lipid bilayer, making helix-helix contacts with the δ and ϵ subunits thus binding F₁ to the membrane. All three F₀ subunits are required to make a transmembrane proton channel (dotted line) linking the membrane proton potential to ATP synthesis in F₁. The genes encoding the ATP-synthase subunits are organized in the *unc* operon as shown in the lower panel. The order of genes was determined by DNA sequence analysis^{16,19,20}. Gene a was identified by homology with mitochondrial ATPase-6 genes²¹, c from total amino acid sequence²²; b and δ from their molecular weight and charge¹⁶ and N-terminal sequences²; α , γ , β and ϵ by sequence analysis of the *E. coli* proteins²³ and, in the case of α and β also from highly homologous beef proteins (M. J. Runswick, V. L. J. Tybulewicz and J.W., unpublished work). The presence of an active promoter preceding gene 1 and the transcription start point have been established by experiments *in vitro* (N.J.G., unpublished work). Parts of the sequence corresponding to genes a, c, b, δ and α have also been sequenced independently^{2,24–26}. Gene 1 does not encode a structural protein of ATP-synthase, but may be involved in regulation of assembly¹⁶.

that the interactions of b with F_1 proteins will be predominantly helix-helix.

The binding of F_1 to F_0 is mediated by δ and ϵ subunits⁶⁻⁹. In thermophile PS3 enzyme, which is very similar to the *E. coli* enzyme¹⁰, both are required to bind the $\alpha\beta\gamma$ subcomplex to F_0

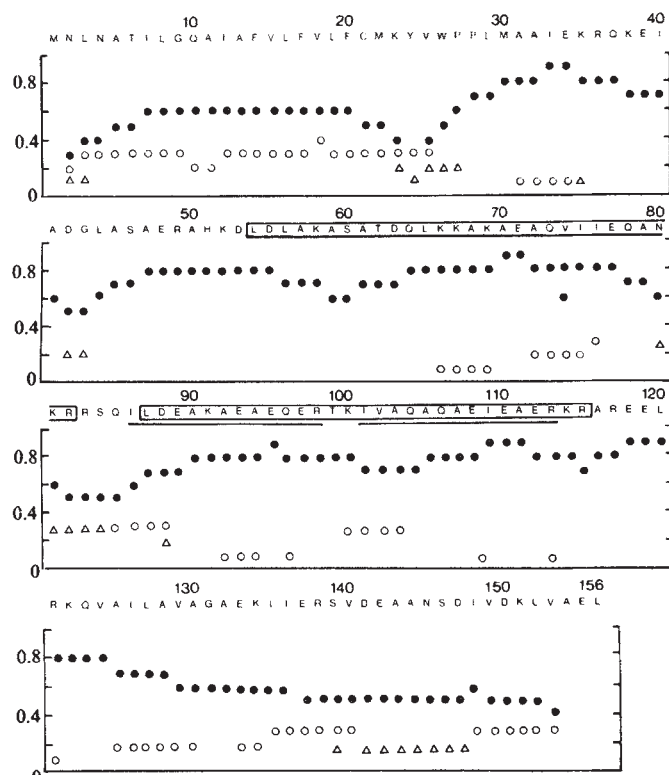


Fig. 2 Amino acid sequence of subunit b, and its secondary structure, predicted from the primary sequence using a program described by McLachlan²⁷. The graphs beneath the sequence show probabilities (ordinate scale) of residues forming an α -helix (●), β -sheet (○) or β -bend (△). Boxed sequences represent related sequences (the long repeat); underlined segments are also related to each other (the short repeat, see Fig. 4).

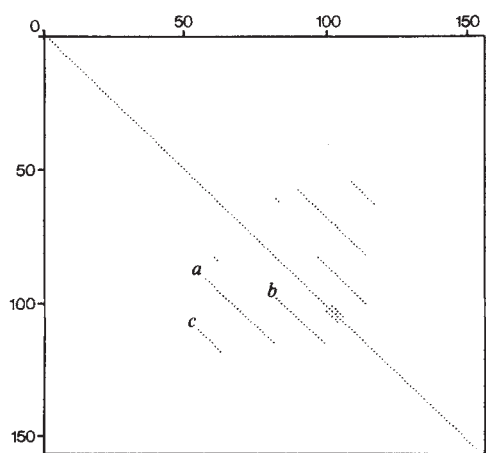


Fig. 3 Matrix of comparison of b with itself, performed with the interactive computer program DIAGON (R. Staden, unpublished work) adapted from earlier procedures²⁸. Scoring is based on a mutation matrix calculated from observed changes in homologous proteins. The scoring level in the matrix shown corresponds to a double matching probability of 1.2×10^{-3} , that is, in a random sequence of the same composition only about one sequence in a thousand should reach this scoring level. The span used in the comparison was 25. Each point (midpoints of the span) has a score less than the 'cutting level'. The diagonal is b aligned with itself; a and b correspond to the long and short repeats, respectively (see Fig. 4), and c corresponds to alignment of residues 57-61 (A K A S A) with residues 103-107 (A Q A Q A), which is probably trivial.

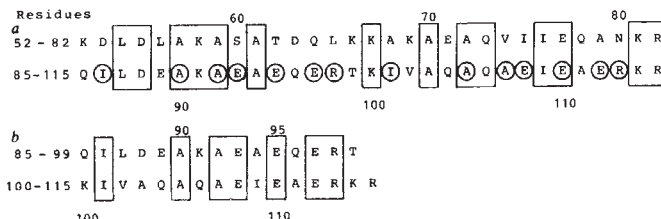


Fig. 4 Alignment of repeated sequences in protein b. a, The long repeats; b, the short repeats. Identities are boxed. In a, the 14 residues conserved in b are encircled; of these four residues are conserved in both a and b.

(ref. 7). Smith and Sternweis showed also that in *E. coli*, δ is involved in attachment of F_1 to F_0 (ref. 7). The δ protein in particular could make helix-helix interactions with b; it has a high α -helical content (65%) as estimated by circular dichroism spectroscopy of the PS3 protein¹¹. In the same study ϵ and γ were estimated to contain 33% and 49% α -helix respectively, in general agreement with α -helix contents of 83%, 31% and 67% for δ , ϵ and γ , respectively, derived from secondary structures predicted from the *E. coli* protein sequences (J.W. and M.S., unpublished work).

The stoichiometry of subunits in F_1 is $3\alpha:3\beta:1\gamma:1\delta:1\epsilon$ ¹² with an estimated one or two copies of b per enzyme complex¹³. Thus it remains unclear whether two or four roughly equivalent α -helical segments in b are available in each F_0 to bind F_1 , also how many complementary α -helices would be required in δ and ϵ .

As it occupies a central position providing an important structural link between the membrane and F_1 , it seems likely that b will be involved directly in proton translocation with some of its charged residues contributing to the process. This is supported by the finding that mutants in b have drastically reduced proton translocation¹. The γ -subunit has also been proposed to participate in this process by gating the proton channel¹⁴. Like protein b it appears to interact with δ and ϵ although not directly with F_0 ^{9,14} and therefore not with b.

Biosynthetic assembly of ATP-synthase is poorly understood. It is clearly desirable to assemble the enzyme in a controlled manner to avoid open H^+ channels. Open H^+ channels would 'de-energize' the membrane leading to breakdown of several essential functions driven by the membrane proton potential¹⁵. The location of b in the enzyme complex, the *in vitro* assembly of F_1 from subunits and reassociation of F_1 and F_0 suggest that b will be important here also. In addition, the order of genes in the *unc* operon (Fig. 1) is striking: the genes for the hydrophobic components of F_0 (a and c) are followed by the gene for protein b, and then those for F_1 , suggesting that insertion of F_0 polypeptides may occur early in assembly. This notion is not in accord with a study of *unc* mutants by Cox *et al.*¹⁵, who proposed that α and β must bind to membranes containing a and c before b can enter the lipid bilayer to complete F_0 (although b was not shown to be present in the cytoplasm of these mutants). However, these studies did not take account of gene 1 (Fig. 1) which may encode a pilot protein to regulate the assembly process¹⁶. Indeed, identification of the role of this gene may be the key to understanding assembly of bacterial ATP-synthase.

Recently Hoppe *et al.*¹⁷ have reported that *E. coli* protein b in F_1 -depleted membranes is very sensitive to proteolytic degradation. Also F_1 protects b in these membranes from proteolysis. Additionally a photoactive hydrophobic phospholipid derivatives reacts with Cys-21. These experiments strongly support the structure of b proposed above.

We thank A. D. McLachlan and R. Staden for help with computer analyses and R. Hendersen and J. Deatherage for helpful comments. M.S. is supported by an EMBO fellowship and N.J.G. by an MRC studentship.

Received 29 March; accepted 22 June 1982.

1. Friedl, P., Bienhaus, G., Hoppe, J. & Schairer, H. U. *Proc. natn. Acad. Sci. U.S.A.* **78**, 6643-6646 (1981).

2. Nielsen, J., Hansen, F. G., Hoppe, J., Friedl, P. & Meyenberg, K. *Molec. gen. Genet.* **184**, 33–39 (1981).
3. Engleman, D. M., Hendersen, R., McLachlan, A. D. & Wallace, B. A. *Proc. natn. Acad. Sci. U.S.A.* **77**, 2023–2027 (1980).
4. Chothia, C., Levitt, M. & Richardson, D. J. *molec. Biol.* **145**, 215–250 (1981).
5. Janin, J. & Chothia, C. *J. molec. Biol.* **143**, 95–128 (1980).
6. Futai, M., Sternweis, P. C. & Heppel, L. A. *Proc. natn. Acad. Sci. U.S.A.* **71**, 2725–2729 (1974).
7. Smith, J. B. & Sternweis, P. C. *Biochemistry* **16**, 306–311 (1977).
8. Yoshida, M., Okamoto, H., Sone, N., Hirata, H. & Kagawa, Y. *Proc. natn. Acad. Sci. U.S.A.* **74**, 936–940 (1977).
9. Sternweis, P. C. *J. biol. Chem.* **253**, 3123–3128 (1978).
10. Takeda, K. *et al. J. Biochem., Tokyo* **91**, 695–701 (1982).
11. Yoshida, M., Sone, N., Hirata, H., Kagawa, Y. & Ui, N. *J. biol. Chem.* **254**, 9525–9533 (1979).
12. Bragg, P. D. & Hou, C. *Archs Biochem. Biophys.* **167**, 311–321 (1975).
13. Fillingame, R. H. *Curr. Topics Bioenerget.* **11**, 35–106 (1981).
14. Sone, N., Yoshida, M., Hirata, H. & Kagawa, Y. *Proc. natn. Acad. Sci. U.S.A.* **75**, 4219–4223 (1978).
15. Cox, G. B. *et al. J. Bact.* **148**, 30–42 (1981).
16. Gay, N. J. & Walker, J. E. *Nucleic Acids Res.* **9**, 3919–3926 (1981).
17. Hoppe, J. *et al. Eur. Bioenerget. Conf. Rep.* **2**, 85–86 (1982).
18. Futai, M. & Kanazawa, H. *Curr. Topics Bioenerget.* **10**, 181–215 (1980).
19. Gay, N. J. & Walker, J. E. *Nucleic Acids Res.* **9**, 2187–2194 (1981).
20. Saraste, M., Gay, N. J., Eberle, A., Runswick, M. J. & Walker, J. E. *Nucleic Acids Res.* **9**, 5287–5296 (1981).
21. Anderson, S. *et al. Nature* **290**, 457–465 (1981).
22. Hoppe, J., Schairer, H. U. & Seibald, W. *Eur. J. Biochem.* **112**, 17–24 (1980).
23. Walker, J. E. *et al. Eur. J. Biochem.* **123**, 253–260 (1982).
24. Kanazawa, H., Mabuchi, K., Kayano, T., Tamura, F. & Futai, M. *Biochem. biophys. Res. Commun.* **100**, 219–225 (1981).
25. Mabuchi, K., Kanazawa, H., Kayano, T. & Futai, M. *Biochem. biophys. Res. Commun.* **102**, 172–179 (1981).
26. Tamura, F., Kanazawa, H., Tsuchiya, T. & Futai, M. *FEBS Lett.* **127**, 48–52 (1981).
27. McLachlan, A. D. *Int. J. Quantum Chem.* **12**, Suppl. 1, 371–385 (1977).
28. McLachlan, A. D. *J. molec. Biol.* **61**, 409–524 (1971).

Evidence for hydrogen bonding of bound dioxygen to the distal histidine of oxycobalt myoglobin and haemoglobin

T. Kitagawa,* M. R. Ondrias† & D. L. Rousseau

Bell Laboratories, Murray Hill, New Jersey 07974, USA

M. Ikeda-Saito & T. Yonetani

Department of Biochemistry and Biophysics,
University of Pennsylvania School of Medicine,
Philadelphia, Pennsylvania 19104, USA

The origin of the differences in oxygen binding energy in various haemoglobins and myoglobins has long been debated. Perutz¹ proposed that the haem-coordinated histidine (proximal histidine) strains the haem iron in low affinity globins but relaxes it in high affinity globins. The existence of such tension in T-structure deoxyhaemoglobin (deoxyHb) was recently confirmed by electron paramagnetic resonance (EPR)^{2,3}, resonance Raman^{4,5} and NMR⁶ spectroscopy. Although its contribution to the free energy of cooperativity is insignificant in the deoxy state, the tension at the haem is considered to be ~ 1 kcal mol⁻¹ for the ligated form in which the haem iron moves into the porphyrin plane⁷. The remaining free energy is probably stored in other parts of the molecule. Therefore, a study of the stabilization mechanisms of the oxygenated form became increasingly important. A hydrogen bond between the bound oxygen and the distal histidine has been proposed by Pauling⁸; this would be expected to stabilize the oxy form of the protein and could contribute to the regulation of the oxygen affinity through the oxygen dissociation rate. A series of EPR and functional studies on various cobalt-substituted monomeric haemoglobins and myoglobins suggested the presence of such hydrogen bonding^{8–12} and it has recently been established in crystals of oxy iron myoglobin (oxyFeMb)¹³ and in oxyhaemoglobin¹⁴. Here we present resonance Raman spectra of the oxy forms of cobalt-porphyrin-substituted myoglobin and haemoglobin (CoMb and CoHb) recorded in buffered H₂O and D₂O solutions at 406.7 nm excitation. Only the Raman lines corresponding to the O—O stretching mode of the bound oxygen¹⁵,

appearing near 1,130 cm⁻¹, are shifted (2–5 cm⁻¹) on replacement of H₂O by D₂O; no other vibrations, including the Co—O₂ stretching mode, exhibit any frequency shifts. This indicates that the bound oxygen in oxyCoMb and in both subunits of oxyCoHb interacts with the adjacent exchangeable proton, and confirms the formation of a hydrogen bond between the bound oxygen and the distal histidine⁹.

The IR active vibrational frequencies of the bound dioxygen stretching modes in haemoglobin have been detected in the 1,100–1,160 cm⁻¹ frequency range¹⁶. Resonance Raman spectroscopy probes the vibrational frequencies of the porphyrin chromophore and electronically associated ligands. Tsubaki and Yu¹⁵ demonstrated by ¹⁶O₂–¹⁸O₂ isotopic substitutions, that the dioxygen stretching mode of the bound oxygen of oxyCoMb and oxyCoHb, located near 1,130 cm⁻¹, was resonance-enhanced on excitation at 406.7 nm (ref. 15). In contrast, the spectra of oxyFeMb and oxyFeHb revealed no oxygen isotope-dependent frequencies in this region. Tsubaki and Yu¹⁵ proposed that the dioxygen stretching mode is enhanced by coupling to a charge transfer transition in the cobalt-substituted haem that is absent from the iron haem. The O—O stretching frequency should be a sensitive probe of hydrogen bonding to the dioxygen molecule, therefore we investigated the frequency of the O—O stretching mode of oxyCoMb and oxyCoHb on deuterium substitution of the exchangeable protons.

CoMb, CoHb and iron-cobalt hybrid haemoglobins ($\alpha(\text{Fe})_2\beta(\text{Co})_2$ and $\alpha(\text{Co})_2\beta(\text{Fe})_2$) were prepared as described previously^{2,17} and oxygenated in 50 mM Tris-HCl buffer pH (or pD) 9.0, except for the pH variation studies. The Raman spectra were recorded using the difference spectrometer described previously¹⁸, the accuracy of the detected difference in the frequency was ~ 0.2 cm⁻¹ and that of the absolute frequency $\sim \pm 1$ cm⁻¹.

Typical H₂O–D₂O difference spectra for oxyCoHb and oxyCoMb in the 1,000–1,250 cm⁻¹ region are shown in Fig. 1a, b. The spectrum of oxyCoHb in H₂O (Fig. 1b) is in agreement with the reported spectrum in which Raman lines near 1,133 and 1,146 cm⁻¹ were found to be ¹⁸O₂ isotope-sensitive¹⁵. These two lines show no frequency shift between pH 5.3 and 10.4. On replacing the H₂O by D₂O, however, the 1,133 cm⁻¹ component exhibits a frequency shift of 5 cm⁻¹ to 1,138 cm⁻¹ (Fig. 1b). There may be a shift in the 1,146 cm⁻¹ line also (see the difference spectrum) but its magnitude is obscured by the shift of the more intense line. In contrast to this behaviour in the cobalt haemoglobin, no frequency differences were detected on deuteration of iron haemoglobin. This was expected as the only Raman-active modes in the 1,050–1,150 cm⁻¹ region of oxyFeHb, are porphyrin modes. Thus, the hydrogen bonding is probably present in the iron haems also but it is not spectroscopically detectable in our conditions. Figure 1a shows the spectra of oxyCoMb in which the H₂O–D₂O shift is from 1,134 to 1,136 cm⁻¹. In oxyCoMb the higher frequency component ($\sim 1,146$ cm⁻¹) has been shown to be weaker than in oxyCoHb¹⁵, thus obscuring any possible shift in frequency in this line.

The iron-cobalt hybrid haemoglobins were also examined. The absence of any H₂O–D₂O dependence in oxyFeHb allows determination of the relative contributions from the individual subunits in the hybrids as any shift must result from the cobalt subunit. The size of the shift and shape of the line on deuteration in the α - and β -subunits are depicted in Fig. 1c, d. The α -subunit gave a smaller shift (2 cm⁻¹) than the β -subunits (4 cm⁻¹). The magnitude of shift for each subunit was less than that of fully cobalt-substituted oxyHb. This is expected as the presence of Raman modes from the iron subunits of the hybrids that do not shift in this frequency region should dilute the contribution from the cobalt subunits. The frequency difference between the $\alpha(\text{Fe})_2\beta(\text{Co})_2$ and $\alpha(\text{Co})_2\beta(\text{Fe})_2$ oxyHbs, both in H₂O, is small. Therefore the frequency difference that we observe between the α - and β -subunits occurs in the deuterated form. A shape difference in this mode between the subunits is also noted in the deuterated form; the 1,136 cm⁻¹ line of the

* Permanent address: Department of Molecular and Physiological Chemistry, Osaka University Medical School, 2-7 Daigakumachi, Osaka, Japan.

† Present address: Department of Chemistry, University of New Mexico, Albuquerque, New Mexico 87131.

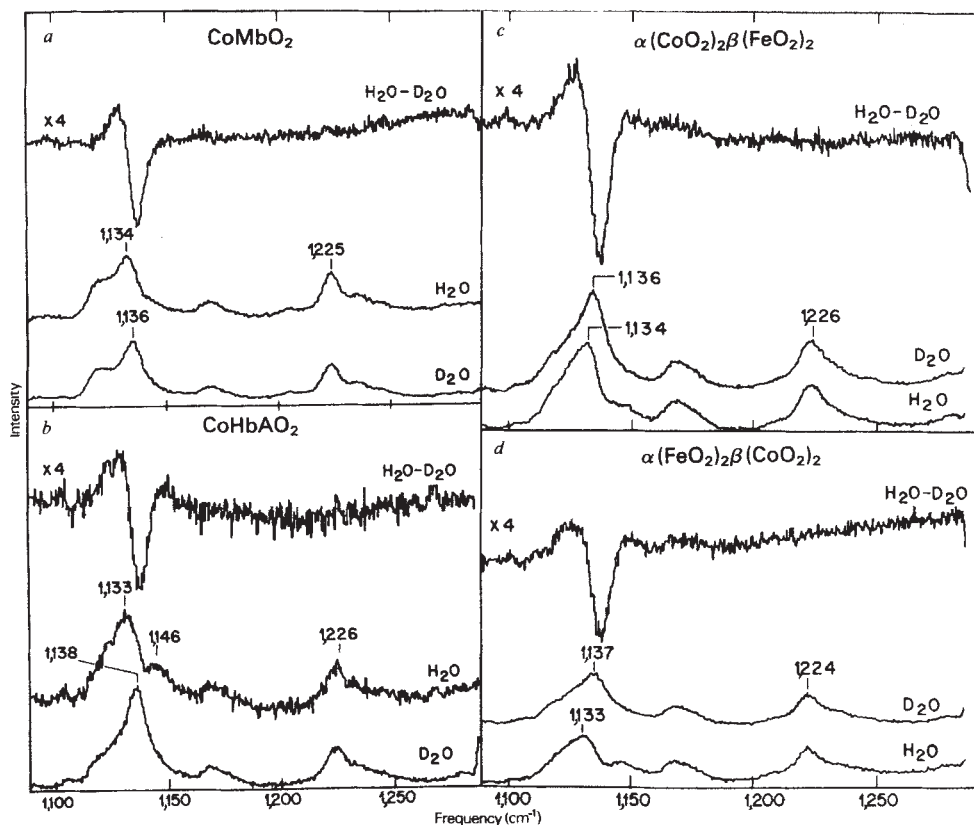


Fig. 1 Resonance Raman spectra of a, oxyCoMb; b, oxyCoHb; c, oxy $\alpha(\text{CoO}_2)_2\beta(\text{FeO}_2)_2$; and d, oxy $\alpha(\text{FeO}_2)_2\beta(\text{CoO}_2)_2$ in H_2O and D_2O . In each case the difference spectrum is plotted at the top with a sensitivity four times as high as the individual spectra. Excitation frequency 406.7 nm, laser power ~ 100 mW, spectral slit width ~ 4 cm^{-1} .

$\alpha(\text{CoO}_2)_2\beta(\text{FeO}_2)_2$ oxygenated hybrid is narrower than that of the complementary hybrid.

Figure 2 shows the $\text{H}_2\text{O}-\text{D}_2\text{O}$ difference spectrum of oxy-CoMb in the $\text{Co}-\text{O}_2$ stretching region. The Raman line of oxyCoMb at 538 cm^{-1} was assigned previously to the $\text{Co}-\text{O}_2$ stretching mode¹⁵. Unexpectedly, this line did not show any frequency difference ($< 2\text{ cm}^{-1}$) in the H_2O and D_2O solutions despite the clear shift of the $\text{O}-\text{O}$ stretching mode.

Deuterium substitution of exchangeable protons of haemoglobin has been reported to stabilize the protein against thermal denaturation and auto-oxidation¹⁹. Thus, the hydrogen bonding in haemoglobin is presumably stronger in D_2O than in H_2O as expected on the basis of a lower zero point energy. Therefore, a resulting change of tertiary structure due to the deuterium substitution could alter a state of the $\text{Co}-\text{His}$ (proximal) bond or a degree of charge delocalization between the haem and the bound dioxygen. This seems unlikely, however, because the $\text{Fe}-\text{His}$ stretching mode of deoxyFeHb is not shifted by more than 1 cm^{-1} on deuterium substitution; nor is any frequency shift detected on deuteration for any of the porphyrin skeletal vibrations of oxyCoHb, and the $\text{Co}-\text{O}_2$ mode is unaffected by deuteration. In addition, Phillips and Schoenborn, using neutron diffraction¹³, observed no significant difference between the tertiary structures of myoglobin in H_2O and D_2O . Therefore, the differences we have detected in the $\text{O}-\text{O}$ stretching frequency in D_2O must result from a specific interaction between the oxygen and the apoprotein on the distal side of the haem.

Recently, Phillips and Schoenborn¹³ made a direct observation of a hydrogen bond in crystals of oxyFeMb and reported that the $\text{N}_\text{e}-\text{O}^2$ distance is $2.97\text{ \AA}$. According to a recent X-ray analysis¹⁴ of crystalline oxyFeHb, the $\text{N}_\text{e}-\text{O}^2$ distance is $2.7\text{ \AA}$ and $3.4\text{ \AA}$ for the α - and β -subunits, respectively. Furthermore, the $\text{N}_\text{e}-\text{O}^2$ distance in crystals of oxyCoMb has also been shown to be compatible with hydrogen bonding²⁰. The data reported here, obtained in solution, confirm the existence of hydrogen bonds in oxyCoMb and in both subunits of oxyCoHb. It is probable that the entire $\text{H}_2\text{O}-\text{D}_2\text{O}$ shift of the $\text{O}-\text{O}$ stretching frequency is caused by the change from the hydrogen-oxygen bond to the deuterium-oxygen bond. It is also

possible, however, that a change in position of the distal histidine due to tertiary structural changes in the protein contributes to our observed frequency differences. The hydrogen bond to the distal histidine is (approximately) perpendicular to the $\text{O}-\text{O}$ bond axis¹³, therefore it seems reasonable that the $\text{Co}-\text{O}_2$ stretching frequency is less sensitive to the changes caused by deuteration than the $\text{O}-\text{O}$ stretching frequency. More information is required to correlate the size and direction

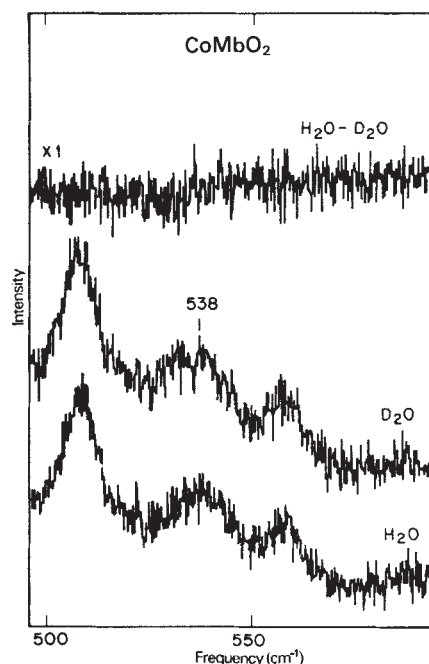


Fig. 2 Resonance Raman spectra of oxyCoMb in the region of the $\text{Co}-\text{O}_2$ stretching frequency (538 cm^{-1}). The absence of a significant shift on deuteration in this mode may be seen by examination of the difference spectrum plotted at the top. Excitation frequency 441.6 nm, laser power ~ 20 mW, spectral slit width $\sim 4\text{ cm}^{-1}$.

of the frequency shift with the strength of the hydrogen bond or with conformational changes. Analysis of the O—O stretching mode in oxyCoHbs in which a quaternary structure change can be induced should elucidate whether any free energy of cooperativity is localized in the hydrogen bonding interaction.

We thank M. F. Perutz for critically reading the manuscript. The portion of the work done at the University of Pennsylvania was supported by National Heart, Lung and Blood Institute grant HL-14508 and NSF grant PCM79-22841.

Received 31 March; accepted 24 June 1982.

1. Perutz, M. F. *Nature* **228**, 726–734 (1970).
2. Ikeda-Saito, M., Yamamoto, H. & Yonetani, T. *J. biol. Chem.* **252**, 8639–8644 (1977).
3. Inubushi, T., Ikeda-Saito, M. & Yonetani, T. in *Hemoglobin and Oxygen Binding* (ed. Ho, C.) 199–204 (Elsevier, New York, 1982).
4. Nagai, K. & Kitagawa, T. *Proc. natn. Acad. Sci. U.S.A.* **77**, 2033–2037 (1980).
5. Ondrias, M. R., Rousseau, D. L. & Simon, S. R. *Science* **213**, 657–659 (1981).
6. Nagai, K. & LaMar, G. N. *Biochemistry* (in the press).
7. Perutz, M. F. *et al. Biochemistry* **17**, 3640–3661 (1978).
8. Pauling, L. *Nature* **203**, 182–183 (1964).
9. Yonetani, T., Yamamoto, H. & Iizuka, T. *J. biol. Chem.* **249**, 2168–2174 (1974).
10. Ikeda-Saito, M., Iizuka, T., Yamamoto, H., Kayne, T. J. & Yonetani, T. *J. biol. Chem.* **252**, 4882–4887 (1977).
11. Ikeda-Saito, M., Brunori, M. & Yonetani, T. *Biochim. biophys. Acta* **533**, 173–180 (1978).
12. Ikeda-Saito, M., Hori, H., Inubushi, T. & Yonetani, T. *J. biol. Chem.* **256**, 10267–10271 (1981).
13. Phillips, S. V. E. & Schoenborn, B. P. *Nature* **292**, 81–82 (1981).
14. Schaanan, B. *Nature* **296**, 683–684 (1982).
15. Tsubaki, M. & Yu, N. T. *Proc. natn. Acad. Sci. U.S.A.* **78**, 3581–3585 (1981).
16. Barlow, C. H., Maxwell, J. C., Wallace, W. J. & Caughey, W. S. *Biochem. biophys. Res. Commun.* **55**, 91–95 (1973).
17. Yonetani, T., Yamamoto, H. & Woodrow, G. V. *J. biol. Chem.* **249**, 682–690 (1974).
18. Rousseau, D. L. *J. Raman Spectrosc.* **10**, 94–99 (1981).
19. Tomita, S. & Riggs, A. *J. biol. Chem.* **245**, 3104–3109 (1970).
20. Petsko, G. A., Rose, D., Tsernoglou, D., Ikeda-Saito, M. & Yonetani, T. in *Frontiers of Biological Energetics* (eds Dutton, P. L., Leigh, J. S. & Scarpa, A.) 1011–1016 (Academic, New York, 1978).

Myosin phosphorylation, agonist concentration and contraction of tracheal smooth muscle

Primal de Lanerolle, John R. Condit Jr, Mark Tanenbaum & Robert S. Adelstein

Laboratory of Molecular Cardiology, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, Maryland 20205, USA

Myosin phosphorylation plays an important part in excitation-contraction coupling in smooth muscle. Phosphorylation by a Ca^{2+} , calmodulin-dependent kinase stimulates the actin-activated Mg^{2+} -ATPase activity of smooth muscle myosin^{1–5}, suggesting that myosin phosphorylation regulates smooth muscle contraction (see refs 5, 6 for discussions of alternative hypotheses). This hypothesis is supported by evidence that myosin is phosphorylated during contraction and dephosphorylated during relaxation of intact smooth muscles stimulated with a single agonist concentration^{7–9}. However, there is little information regarding the response to stimulation with various agonist concentrations. As the dose-response relationships for phosphorylation and tension should be similar if myosin phosphorylation does, in fact, regulate smooth muscle contraction, we studied myosin phosphorylation in tracheal smooth muscle stimulated with a broad range of concentrations of the cholinergic agonist, methacholine. The results of these experiments are consistent with the hypothesis that myosin phosphorylation regulates smooth muscle contraction but they indicate a relatively complex relationship between myosin phosphorylation and the generation of isometric tension.

Canine tracheal smooth muscles were obtained and treated as previously described⁸ except that: (1) muscles were maintained in the muscle baths as flat sheets in order to minimize non-uniform rates of agonist diffusion to receptor sites; (2) individual smooth muscle strips were maintained at a length of 2 cm, as defined by their *in vivo* length, from the time they were removed from the cartilaginous rings of the trachea until the end of the experiment; (3) muscle strips were kept in the muscle baths for 45 min before initiating contractile responses.

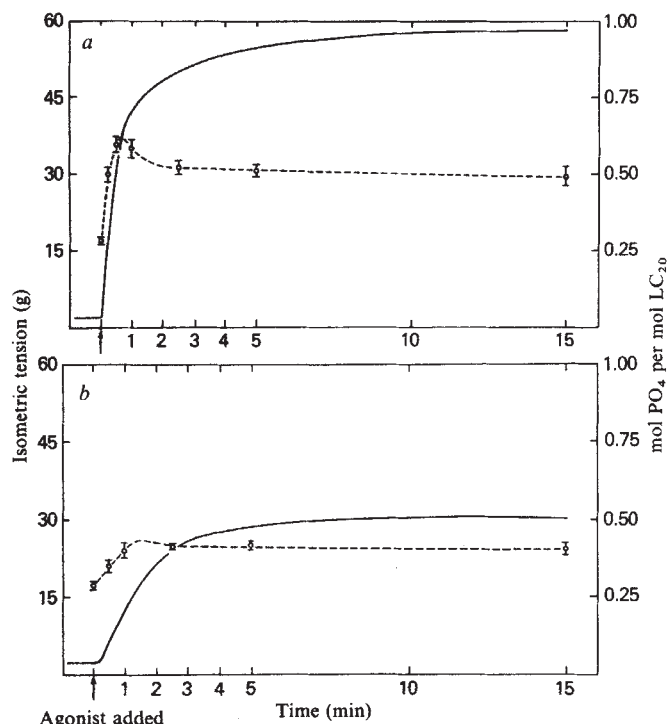


Fig. 1 Myosin phosphorylation and tension generated by tracheal smooth muscle contracted with 10^{-4} M (a) or 2×10^{-7} M (b) methacholine. Tracheal muscle from six consecutive tracheal rings was isolated and mounted in six myobaths. The data at 15 min in a were obtained from separate experiments. The solid lines represent tension generated by representative muscles contracted for 15 min; broken lines represent the phosphate content of muscles frozen at the times indicated. Error bars indicate the s.e.m. for phosphate determinations on at least five different pieces of tissue. Experiments on smooth muscles have demonstrated that a specific subunit of the myosin molecule, with a molecular weight of 20,000 (LC_{20}), is phosphorylated. A maximum of 1 mol PO_4 can be incorporated per mol LC_{20} . As there are 2 mol LC_{20} per mol myosin, the maximal level of myosin phosphorylation is 2 mol PO_4 per mol myosin. We have chosen to present the data as mol PO_4 per mol LC_{20} because this was the experimentally determined value.

The myosin phosphate content was quantitated using an immunoprecipitation–isoelectric focusing method as described previously⁸.

Phosphate analyses on resting (unstimulated) tracheal muscle demonstrated 0.29 mol of phosphate incorporated per mol LC_{20} . (LC_{20} is a subunit of myosin—see Fig. 1 legend.) On stimulation with 10^{-4} M methacholine (Fig. 1a), the myosin phosphate content rose rapidly, reaching a peak of 0.62 mol PO_4 per mol LC_{20} , before the muscle generated maximal tension. The myosin phosphate content then decreased slightly, while the muscle was still contracting, to a stable level of 0.49 mol PO_4 per mol LC_{20} . Stimulation of tracheal muscle with 2×10^{-7} M methacholine (Fig. 1b), a concentration previously determined to produce half-maximal tension, contracted muscles more slowly than stimulation with 10^{-4} M methacholine (maximal rates of contraction were 31 and 129 g min^{-1} , respectively). Myosin was phosphorylated at a much slower rate, and the maximal and steady-state levels of phosphorylation were lower for muscles stimulated with 2×10^{-7} M methacholine (Fig. 1b) than for those stimulated with 10^{-4} M methacholine (Fig. 1a).

We have also determined the relationship between the steady-state level of myosin phosphorylation, tension and agonist concentration (Fig. 2). Muscles were stimulated with various concentrations of methacholine, then frozen 15 min after addition of the agonist and assayed to determine the myosin phosphate content. (Muscles were frozen at 15 min because the experiments in Fig. 1 indicated that tension and phosphorylation had reached a plateau by 10 min.) Myosin phosphorylation and active tension were expressed as a percentage of the

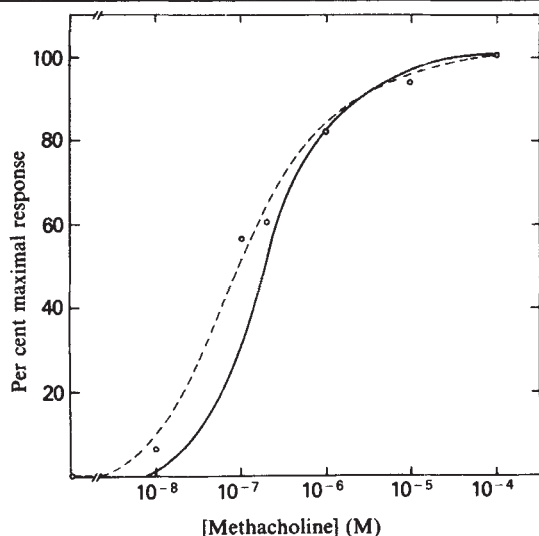


Fig. 2 The dose-response relationship for myosin phosphorylation and isometric tension in tracheal smooth muscle stimulated with methacholine. Myosin phosphorylation was determined on tracheal smooth muscle contracted for 15 min using the agonist concentration shown. The phosphorylation and tension data were expressed as per cent maximal increase using the equation $X - R/M - R$ where X = myosin phosphate content or tension at 15 min of muscles stimulated with various methacholine concentrations; R = the resting phosphate content (29%) or resting tension; and M = the myosin phosphate content (49%) or tension at 15 min of muscles stimulated with 10^{-4} M methacholine. The solid line represents the rise in tension after stimulation at each agonist concentration; the broken line represents the increase in myosin phosphorylation at each agonist concentration. Each point represents the mean from phosphate assays on at least five different pieces of tissue.

maximum (see Fig. 2 legend). These data, plotted as a function of log agonist concentration (Fig. 2), demonstrated similar agonist concentration-response relationships for both phosphorylation and tension. Analysis of variance indicated statistically significant differences ($P < 0.001$) between successive points on each dose-response curve. However, there were no statistically significant differences between the two curves at any given agonist concentration, indicating that the phosphorylation and tension data define superposable curves.

The data on myosin phosphorylation presented here argue against a simple relationship between the processes of myosin phosphorylation and tension generation by tracheal smooth muscle. Three major aspects of the phosphorylation data require comment—the resting myosin phosphate content, the overshoot in myosin phosphorylation and the steady-state level of myosin phosphorylation. Each of these may have a bearing on the manner in which tracheal muscle generates tension and will be discussed separately.

First, it should be noted that a substantial fraction of the myosin (29%) is phosphorylated in unstimulated tracheal smooth muscle. There is a low level of tension (2–5 g) in these muscles but this resting tension seems to be generated by passive elastic elements as a reduction in the myosin phosphate content does not decrease the resting tension⁸. It is possible that a high level of myosin phosphorylation is unique to tracheal smooth muscle because lower levels of myosin phosphorylation have been reported in resting^{10,11} and pharmacologically relaxed¹² smooth muscles. Nevertheless, one explanation for this observation is that these represent monophosphorylated myosin molecules. Recent experiments¹³ on the relationship between smooth muscle myosin phosphorylation and acto-myosin ATPase activity indicate the apparent requirement that both myosin heads be phosphorylated in order for the myosin to be activated by actin. Thus, it is possible that myosin molecules involved in generating tension in tracheal smooth muscle are monophosphorylated in resting conditions and do not begin to hydrolyse ATP and generate tension until the second head is phosphorylated.

Second, what is the significance of the overshoot in myosin phosphorylation seen in muscles maximally stimulated with methacholine (Fig. 1a)? Recent studies on intracellular calcium levels during contraction of smooth muscle cells^{14,15} and stimulation of non-muscle cells¹⁶ have demonstrated that intracellular calcium levels rise sharply, peak and then decay after stimulation of these cells. Furthermore, peak levels of intracellular calcium are reached before the smooth muscle cells generate maximal tension. As myosin kinase requires Ca^{2+} -calmodulin for activity³, the activation of this enzyme probably follows the calcium transient. In fact, the overshoot in myosin phosphorylation may be a reflection of the redistribution of the intracellular calcium levels in response to stimulation of muscarinic receptors¹⁷.

Finally, the most direct relationship between myosin phosphorylation and tension appears to be between the steady-state level of myosin phosphorylation and maximal isometric tension generated at a given agonist concentration. The data in Fig. 2 support this hypothesis as they demonstrate that both myosin phosphorylation and isometric tension are coupled to stimulation of muscarinic receptors. Uncoupling of either process, or a shift in the phosphorylation curve in relation to the tension curve, would have indicated a complex relationship between tension and steady-state myosin phosphorylation. The generation of superposable dose-response curves for phosphorylation and tension suggests that the active isometric tension maintained by tracheal smooth muscle is directly related to the steady-state level of myosin phosphorylation. The data also suggest that the myosin phosphorylation-dephosphorylation reaction can be regulated in tracheal smooth muscle so as to maintain defined levels of phosphorylation which, in turn, seem to determine the tension generated by the muscle.

In contrast to the data presented above, myosin phosphorylation does not remain elevated in hog carotid artery smooth muscle contracted with KCl¹⁰. In this case the myosin phosphate content decreased to a level slightly above the resting level after the carotid artery muscle had generated maximal tension, which suggested that smooth muscles can maintain tension through the formation of a special, non-cycling acto-myosin cross-bridge, termed a 'latch' bridge¹⁰. It is possible that latch bridges are formed only when carotid artery smooth muscles are stimulated by KCl, as the myosin phosphate content remains elevated in the same muscles contracted with histamine¹¹. Although the presence of latch bridges in smooth muscles contracted with pharmacological agents cannot be discounted, the demonstration that the rise in both tension and myosin phosphorylation is related to the level of agonist stimulation (Fig. 2), and the observation that myosin phosphorylation remains elevated when smooth muscles are stimulated with pharmacological agonists (Fig. 1; and ref. 11), suggest that the steady-state level of myosin phosphorylation may be the major determinant of the tension maintained by a smooth muscle.

Thus, the data presented here support, but do not prove, the hypothesis that myosin phosphorylation regulates muscle contraction, and also demonstrate a relatively complex relationship between myosin phosphorylation and the process of tension generation by tracheal smooth muscle.

Received 13 April; accepted 18 June 1982.

- Gorecka, A., Aksoy, M. O. & Hartshorne, D. J. *Biochem. biophys. Res. Commun.* **71**, 325–331 (1976).
- Sobieszek, A. *Eur. J. Biochem.* **73**, 477–483 (1977).
- Dabrowska, R., Sherry, J. M. F., Aromatorio, D. K. & Hartshorne, D. J. *Biochemistry* **17**, 253–258 (1978).
- Sellers, J. R., Pato, M. D. & Adelstein, R. S. *J. biol. Chem.* **256**, 13137–13142 (1981).
- Chacko, S. & Rosenfeld, A. *Proc. natn. Acad. Sci. U.S.A.* **79**, 292–286 (1982).
- Ebashi, S. *Proc. R. Soc. B207*, 259–296 (1980).
- Barron, J. T., Barany, M., Barany, K. & Storti, R. V. *J. biol. Chem.* **255**, 6238–6244 (1980).
- de Lanerolle, P. & Stull, J. T. *J. biol. Chem.* **255**, 9993–10000 (1980).
- Janis, R. A., Barany, K., Barany, M. & Sarmiento, J. G. *Molec. Physiol.* **1**, 3–11 (1981).
- Dillon, P. F., Aksoy, M. O., Driska, S. P. & Murphy, R. A. *Science* **211**, 495–497 (1981).
- Aksoy, M. O., Murphy, R. A. & Kamm, K. E. *Am. J. Physiol.* **242**, C109–C116 (1982).
- Somlyo, A. V., Butler, T. M., Bond, M. & Somlyo, A. P. *Nature* **294**, 567–569 (1981).
- Persechini, A. & Hartshorne, D. J. *Science* **213**, 1383–1385 (1981).
- Fay, F. S., Shlevin, H. H., Granger, W. C. Jr & Taylor, S. R. *Nature* **280**, 506–508 (1979).
- Neering, I. R. & Morgan, K. G. *Nature* **288**, 585–587 (1980).
- Tsien, R. Y., Pozzan, T. & Rink, T. J. *Nature* **295**, 68–71 (1982).
- Bolton, T. B. *Physiol. Rev.* **59**, 606–718 (1979).

BOOK REVIEWS

But is man a worm?

J. S. Jones

THE most difficult thing in writing a book about evolution is to know what to leave out. Almost any biological observation can be discussed in Darwinian terms: indeed, the very universality of the theory of evolution is the main philosophical obstacle to subjecting it to scientific test. In the past few years we have had books dealing with, for example, evolutionary biochemistry, the evolution of animal behaviour and the mathematics of the evolutionary process, together with a mass of popular (or at least unscientific) works which purport to explain human behaviour and social organization in terms of evolutionary theory. In the rush to explain such complex matters in simple terms there has often been a danger of losing sight of the biological wood through the evolutionary trees. What interested Darwin was, of course, the diversity of living organisms. It is therefore surprising how small a part the variety of animals and plants plays in most of these volumes.

Here we have a book, aimed both at undergraduates and at a general audience, which sets out not only to describe the mechanisms and results of evolutionary change in a range of species from cycads to cephalopods but attempts also to apply such patterns to the evolution of human behaviour and the course of human history. Such attempts — and there are many of them — imply that man is indeed little more than a remarkably well-developed worm; the same rules which govern the evolution of amphibia from fish or of *Homo sapiens* from *Ramapithecus* have been responsible for the change from tribe to nation state or from feudalism to capitalism. How well does G. Ledyard Stebbins, an eminent plant evolutionist, succeed in establishing a seamless transition from the primordial slime to the Reagan administration?

It is immediately obvious that Stebbins is far better qualified to attempt this task than are most of his recent predecessors. Instead of the usual facile discussions of human society in terms of carefully selected anecdotes of animal behaviour, we have in the first two-thirds of the book a greatly compressed but remarkably complete account of modern evolutionary biology.

Darwin to DNA, Molecules to Humanity. By G. Ledyard Stebbins. Pp.491. Hbk ISBN 0-7167-1331-4; pbk ISBN 0-7167-1332-2. (W. H. Freeman: 1982.) Hbk \$25, £21.80; pbk \$12, £10.50.

The results of this compression are sometimes alarming; for example, Chapter 9 takes us "From Worms to Primates" in 31 pages of text. There is nevertheless much more insight into the process of evolution here than in many books claiming a far higher level of technical sophistication.

IMAGE
UNAVAILABLE
FOR
COPYRIGHT
REASONS

Many of these insights are attractively phrased: thus Stebbins illustrates the inevitable incompleteness of the fossil record by pointing out that no one has seen the fossilized remnants of a passenger pigeon, although these birds flourished in millions in North America only a hundred years ago. There is even room for a few poetic touches, as when he talks of the oppressive silence of the Carboniferous forests before the evolution of animals capable of making sounds.

Stebbins illustrates the changes in the fortunes of Darwinism over the past century by recalling disparagingly that in his Harvard days in 1926 the theory outlined in the *Origin of Species* was described as having been long ago abandoned. The recent resurgence of anti-

Darwinian models of speciation by "punctuated equilibrium" from the same institution shows how rapidly the pendulum of evolutionary fashion can swing. Stebbins himself considers that the pattern of change of animals and of societies does indeed show stability followed by relatively rapid change, although he is careful to point out that the enormous time scale of evolution in comparison to the brief generation interval of living organisms means that this can easily be accommodated in the Darwinian model.

Only three years after the publication of the *Origin*, Marx wrote to Engels:

It is remarkable how Darwin recognizes among beasts and plants his English society with its division of labour, competition, opening up of new markets, inventions, and the Malthusian "struggle for existence".

Darwin was himself well aware of the possible application of his theory to human history but steadfastly refused to be drawn into discussion on the subject. Such restraint has scarcely been a characteristic of his successors. Indeed it is now almost an occupational hazard for elderly and distinguished biologists to produce biological overviews of human society. Stebbins is no exception, and it is in the last third of his book — "Human Evolution: Biological and Cultural" — that the clarity of his arguments begins to falter.

Although Stebbins criticizes some of the particularly inane statements of geneticists about history (such as Darlington's equation of the success of the Normans with their acquisition of genes both for horse-riding and for seamanship) he is not inhibited from discussing patterns of historical change — including such unjustifiably neglected episodes as the Revolt of the Red Eyebrows in first-century China — in genetic terms. This does provide occasional brilliant insights into human history. For example, the lack of speciation in the hominid line compared to that in other primates is ascribed to the development of tools, which themselves underwent adaptive radiation. However, for much of the time he is drawn onto the wilder shores of speculation as, for instance, when he discusses the presumed coincidence of nut-eating with the beginnings of intelligence,

or of mental alertness with the idle ruminations of prehistoric night-watchmen by their newly-discovered fires and the favours bestowed on them by females as rewards for their wakefulness.

Stebbins is too good a scientist to draw simple lessons from the faulty and imperfect evidence of human history. As a result, he often reaches conclusions so broad as to be almost lacking in content:

A safe generalization is that genetic templates transmit potentialities and capacities rather than adult behavioural traits . . . In humans, cultural fitness, based to a large extent on a desire for approval by society, has both a genetic and a cultural basis. Is *Homo sapiens* just another species of animal that has evolved in the same way as have the various species of monkeys and apes? The answer to this question is a

qualified 'no'.

In general, a large part of his answer to the question posed throughout — Can the evolution of man and society be explained in terms of the evolutionary rules affecting lower organisms? — is a similarly resounding "perhaps". The book therefore ends in a curiously flatfooted way: "Is the evolutionist better equipped than anyone else to recommend what is best for human society? Perhaps not". This conclusion is an honest one, but betrays the excellence of Stebbins's treatment of the evolutionary process. *Darwin to DNA* is fine; *Molecules to Humanity* could be considerably edited with little loss. □

J.S. Jones is in the Department of Genetics and Biometry at University College London.

Putting the forams in order

John W. Murray

Foraminifera. By John R. Haynes. Pp.433. ISBN 0-333-28681-2. (Macmillan Press/Halsted: 1982.) £50, \$79.95.

FORAMINIFERA are tiny animals that build themselves a test or shell. Their fossil remains occur in astronomical numbers from the Cambrian to the present day in sediments accumulating in estuaries, seas and oceans. These attributes, together with their great diversity, have made them of use to geologists both for correlating rocks of similar age (biostratigraphy) and for determining the environments of deposition of their enclosing sediment (palaeoecology). In addition, as foraminiferal tests survive the process of drilling boreholes, they are especially useful in the exploration for petroleum.

A sound taxonomic basis is an essential prerequisite for all such studies and Dr Haynes has taken this as the main theme of his book. In particular, his thorough account of how to write the description of a species and the compilation of a synonymy will prove invaluable to postgraduate students and researchers.

The main thrust of the volume is the introduction of a revised suprageneric classification. As the study of Foraminifera has progressed over the past 200 years, the importance of certain morphological features as criteria for grouping or dividing the species into genera, superfamilies and so on have varied. The revised classification follows previous ones in using the wall structure of the test as the primary means of grouping and subdivision but it takes into account the new information amassed from electron microscopy. Greater importance is attached to details of the aperture than is customarily the case.

Nine orders and thirty-two superfamilies are described. For each order there is discussion of the range of morphology observed, possible evolutionary relationships, an appraisal of their stratigraphic value, comments on their ecology and palaeoecology, and a summary of the classification. The latter includes diagnoses of the superfamilies and of the important genera.

The book is described in the preface as a general text which lays stress on the stratigraphical application of Foraminifera. This aim it achieves admirably with a synthesis of knowledge from more than 1,000 publications, but, in addition, it contains much original work. All those involved in the study of Foraminifera will find it both instructive and stimulating reading. □

J. W. Murray is Professor of Geology and Head of the Geology Department at the University of Exeter.

Naked protoplasm and synchronous nuclei

Michael Carlile

Cell Biology of Physarum and Didymium. Vol. 1, *Organisms, Nucleus and Cell Cycle*. Edited by Henry C. Aldrich and John W. Daniel. Pp.444. ISBN 0-12-049601-1. (Academic: 1982.) \$55, £36.40.

THE myxomycete plasmodium is a remarkable object. A mass of protoplasm not subdivided into cells it can, in the most intensively studied species *Physarum polycephalum*, attain the size of a dinner plate and contain hundreds of millions of nuclei which divide synchronously, at least twice a day in a well-nourished plasmodium.

A handful of enthusiasts for "naked protoplasm" have studied myxomycete plasmodia for over a century and since 1961, when workers at the McArdle Laboratory for Cancer Research at Madison grew plasmodia of *P. polycephalum* in pure culture on soluble media, many molecular, developmental and cell biologists have carried out work on this organism. A second species, *Didymium iridis*, has also been the subject of ultrastructural and genetical studies, but here biochemical investigation has been limited through failure to grow plasmodia in pure culture.

In spite of the wide interest in *Physarum* plasmodia, however, reviewing activity has been limited. Two recent works are now remedying this deficiency. In 1980 *Growth and Differentiation in Physarum polycephalum*, edited by W. F. Dove and H. P. Rusch, was published by Princeton University Press (for review see *Nature* 288, 307; 1980). Now a different group of authors have contributed to the first of two volumes edited by Henry Aldrich, a leading student of myxomycete ultrastructure, and John Daniel, who with Howard Rusch was responsible for the crucial 1961 paper on pure culture.

The introductory section of the first

volume consists of a chapter on the morphology and taxonomy of myxomycetes by Alexopoulos and one on *D. iridis* by Collins and Betterley, leading authorities on this organism. Subsequent pages are effectively confined to *P. polycephalum*. Tyson deals with periodic phenomena — not only synchronous nuclear division and other periodic events of the nuclear cycle, but also the control of the protoplasmic streaming that occurs in well-defined channels in the plasmodium at rates of up to a millimetre a second with reversals in the direction of flow taking place at approximately one minute intervals. The plasmodium is capable of rapid migration and among the sensory systems guiding migration is chemotaxis, which is being intensively studied in Japan and is reviewed here by Ueda and Kobatake. Protoplasmic streaming and its control is further considered by Kessler.

The difficult subject of *Physarum* genetics is dealt with ably by Jennifer Dee, the founder of the subject. Two chapters by Mohberg and Babcock then clarify in a succinct way the confusing topics of ploidy and of the genealogy of the strains of *P. polycephalum* in current use. The final four chapters of the first volume deal with nuclear organization (Lafontaine and Cadrin), chromosomal organization (Matthews and Bradbury), DNA (Evans) and RNA (Braun and Seebeck), the important topics of differentiation, metabolism and methodology being deferred to Vol. 2.

This first volume is lucidly written and admirably illustrated. I suspect — and hope — that the two-volume work will do much to promote work on myxomycetes and publicize the results among other biologists. □

Michael Carlile is Senior Lecturer in the Department of Pure and Applied Biology, Imperial College, University of London.

Extra genes, in and out of chromosomes

Joseph G. Gall

Gene Amplification. Edited by Robert T. Schimke. Pp.339. ISBN 0-87969-151-4. (Cold Spring Harbor Laboratory: 1982.) \$35 (US), \$42 (elsewhere).

WHEN mammalian tissue cultures are treated with gradually increasing concentrations of methotrexate (MTX), cell lines can be selected that are resistant to very high concentrations of the otherwise toxic drug. The basis of resistance may be decreased permeability of the cells to the drug or decreased binding between the drug and the target enzyme, dihydrofolate reductase (DHFR). In a significant number of lines, however, resistance is due to compensatory production of the enzyme, which itself is related to increased synthesis of the mRNA and ultimately to differential replication or amplification of the gene coding for DHFR.

These surprising facts were first reported four years ago in elegant experiments by Robert Schimke and his colleagues at Stanford University. Progress in the field has been rapid since then, and late last year Schimke organized a small symposium at Cold Spring Harbor to summarize that progress. This volume, *Gene Amplification*, contains 42 short papers based on presentations at the symposium.

Many of the papers are concerned with the chromosomal and molecular state of the amplified genes. It became evident early on that two unusual cytological phenomena could be associated with amplified cell lines: first, the presence of small acentric chromosomes, which, because they often occur in pairs, are called double minutes; and second the appearance of long chromosome segments with little evidence of G-banding, referred to as homogeneously staining regions. Amplified genes have now been demonstrated in both regions by *in situ* nucleic acid hybridization. There is a good but not perfect correlation between stability of the amplified condition (when MTX is removed) and the physical location of the amplified genes — homogeneously staining regions being characteristic of stable lines, double minutes of unstable ones. It is possible, therefore, that amplified gene copies appear first as small extra-chromosomal elements that later become stably integrated into one or more chromosomes. The size of the amplified region has been measured by hybridizing DNA from amplified lines with genomic sequences that flank the DHFR gene. The amplified segment may be up to 100 kb in length, considerably longer than the coding region itself (31 kb counting its five introns).

Almost nothing is known about the initial molecular events of amplification in tissue cultures. A model discussed in the book by Botchan and by Spradling involves multiple rounds of replication at a

single origin of replication, giving rise to an "onion-skin" configuration. An attractive feature of this model is the ease with which it can be applied to the chorion genes of *Drosophila*, whose amplification in the follicle cells of the oocyte has been analysed in detail by Spradling. In this case the chorion genes are more highly amplified than surrounding sequences, the level of amplification falling off progressively on both sides of the genes. This is just what one would expect if multiple replication forks move away from an origin of replication in or near the genes.

Several other specific amplifications have been induced in tissue cultures, of which the best characterized involve resistance to *N*-(phosphonacetyl)-L-aspartate or to the cadmium ion. These cases involve overproduction of the multifunctional protein "CAD" and of metallothionein-1 respectively. Several other likely cases are discussed in the book, and it seems probable that many genes can be amplified if a suitable inhibitor of enzyme activity is available. An interesting possibility is that random genomic fragments regularly detach from chromosomes and require only a selective force to undergo differential replication.

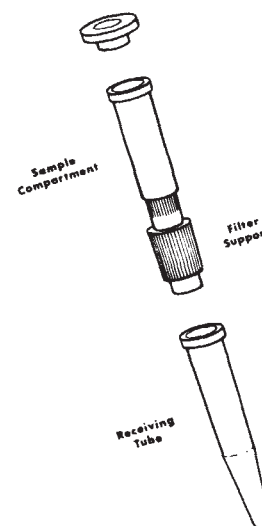
All 42 articles in the volume are short; indeed some are little more than expanded abstracts. The quality is variable, and since many are rather specialized the general reader will have to pick and choose to avoid getting bogged down in cytological and molecular detail. In addition, the title *Gene Amplification* is somewhat misleading. Schimke's major interest is drug-induced gene amplification in cultured cells, and that is what the book really covers. Gene amplification as a general biological phenomenon is touched upon, but readers unfamiliar with the field would not guess that naturally occurring gene amplification, particularly of rDNA, has been investigated extensively during the past 15 years in a variety of vertebrates, insects and protozoa. For a broader introduction to the topic one should begin with Adrian Bird's review article, "Gene reiteration and gene amplification", published in Vol.III of Prescott and Goldstein's *Cell Biology* (Academic, 1980).

Nevertheless this volume summarizes a large amount of work on an important topic. Like the much larger 1980 Cold Spring Harbor Symposium on movable genetic elements, it emphasizes the plasticity of the eukaryotic genome, a subject of great current interest. Because the book was published just six months after the meeting itself, the articles are timely and will prove useful to a wide audience. □

Joseph G. Gall is Professor of Biology, Molecular Biophysics and Biochemistry at Yale University.

Centrifugal Microfilters

Bioanalytical Systems introduces a centrifugal microfilter for membrane filtration of small volumes using the force of a conventional bench top centrifuge. Ideal for LC sample preparation, extraction of TLC spots, and many other applications.



- minimize loss and contamination
- no vacuum or pressure required
- solvent resistant polyalamar
- filter many samples simultaneously
- a variety of filter materials are available
- reusable

Send for details...



**BIOANALYTICAL
SYSTEMS INC.**

1205 Kent Avenue
West Lafayette, IN 47906
(317) 463-2505 telex 276141

Meteoritics as petrography

John A. Wood

Meteorites: A Petrologic-Chemical Synthesis. By Robert T. Dodd. Pp.432. ISBN 0-521-22570-1. (Cambridge University Press: 1982.) £35, \$69.50.

METEORITES are complex, ancient, enigmatic rock fragments from somewhere beyond Mars. They were formed at the same time as the Solar System, and they retain a cryptic record of the events of that time. In the 30 years since Harold Urey and Harrison Brown unleashed the techniques and concepts of post-War geochemistry on these objects, meteoritics has burgeoned into an important scientific subdiscipline. Many studies have been carried out and as many papers published; a number of important insights have been gained; and several riddles are still with us, more prickly and intractable than ever.

The field has always needed a good reference book, the more so as its literature has expanded to a scale greater than any one person can assimilate. Several books about meteorites have been written at a professional level in the past decade, but none can be said to fill the need. Robert Dodd's *Meteorites: A Petrologic-Chemical Synthesis* is the latest candidate. It is important to include the subtitle, because the emphasis of the book is very strongly on the petrology (or, to be more precise, the petrography) and chemistry of meteorites, Dodd's own speciality. (Petrography is the descriptive study of the microstructures and component mineralogies of rocks.)

Many good things can be said about the book. Dodd writes in a readable, straightforward style. His coverage of petrography and chemistry is comprehensive, scholarly and exhaustively referenced. Material is rationally organized, most of the chapters dealing with discrete subclasses of meteorites. (Two exceptions are a chapter on radiometric studies of meteorites, and one on the asteroid-sized parent bodies that they once resided in.) He strikes an appropriate balance between descriptions and interpretive discussions. In the latter, he is as objective as can reasonably be expected of a mortal in this contentious field. Each chapter is preceded by a précis that maps out for the reader the ground to be covered. There is an excellent index, and the illustrations are more than adequate in number and quality.

However, I fault the book for presenting a narrow view of meteoritics, that is, one largely limited to petrographic and geochemical studies. Perhaps it is unreasonable to expect Dodd, or anyone, to cover the whole marvellously diverse field of meteoritics as comprehensively as the author has treated meteoritic petrography and chemistry. The book would be twice as long and would cost Dodd far more than twice the labour, since

he would have to venture beyond his field of experience.

But the book as it stands unfortunately presents the study of meteorites as an end in itself rather than as a means to a greater end: an understanding of the origin of the Solar System. The stage on which this drama occurred was the solar nebula, a giant accretion disk of gas and dust where the properties of the meteorites and, ultimately, the planets, were established. Dodd cannot avoid making repeated references to the nebula, the context in which everything happened; but he eschews any systematic discussion of astrophysical models of the nebula because they are 'varied and changeable'. Lacking such a framework, he cannot put the properties of meteorites to work in clarifying our view of the nebula.

"Nebula" does not appear in the index of the book; nor does "Solar System, origin"; nor does "FUN inclusions". The latter, the subject of intensive isotopic research which gives us a fascinating view of pre-Solar System events, are touched on only momentarily, and then not by the name "FUN". It is regrettable that the non-petrographic aspects of meteorites are given such short shrift: the most inspiring aspects of meteoritics are the grand vista it examines, and the close collaboration of researchers from many diverse disciplines in attacking the problem. *Meteorites: A Petrologic-Chemical Synthesis* exposes only a near-sighted and parochial view of the goals of this field. It is, however, a useful and excellent reference work within the unfortunately restrictive boundaries it has erected. □

John A. Wood is an Associate Director of the Harvard-Smithsonian Center for Astrophysics, Cambridge, Massachusetts.

Welcome to theoretical particle physics

P.V. Landshoff

Gauge Theories in Particle Physics: A Practical Introduction. By Ian J.R. Aitchison and Anthony J.G. Hey. Pp.341. Pbk ISBN 0-85274-534-6. (Adam Hilger/Heyden: 1982.) £14.50, \$36.

THERE are still very few books on the topics that now excite the interest of particle theorists — the electroweak theory that brings together electromagnetism and the weak nuclear interaction, and quantum chromodynamics, the candidate theory of the strong nuclear interaction. So this book, which sets out to be a compact introduction to these subjects, will be warmly welcomed, especially for the attention it gives to the distinctive properties of the gauge symmetry underlying the new theories.

Quantum electrodynamics is the prototype gauge theory and is described by Aitchison and Hey as "the best theory we have". Its best test is the magnetic moment of the electron, whose calculated and measured values agree to within one part in 10^{10} . Electrodynamics is a gauge theory because its physical predictions are invariant under certain transformations of the electromagnetic potentials. Only this gauge invariance makes it possible to remove unphysical infinities from the quantized version of the theory (renormalization).

Quantum electrodynamics is based on Maxwell's equations, the first description of the electric and the magnetic force in the same set of equations. A century later, Glashow, Salam and Weinberg enlarged the theory so as to incorporate the weak interaction (responsible, for example, for neutron β -decay). Their unified electro-

weak gauge theory is steadily accumulating an impressive list of successes.

Another gauge theory, chromodynamics, seems to give a good description of the nuclear force. Attempts to unify this with the electroweak theory have led to the prediction that protons are unstable, with a lifetime of about 10^{31} years. Four candidate proton-decay events have recently been observed, but the data are not yet a sufficient test.

Quantum electrodynamics itself is far from easy, but the electroweak theory and chromodynamics are less accessible to most physicists because the groups of gauge transformations are not Abelian (which implies that the underlying gauge transformations do not necessarily commute with each other). This book sets out to make the theories as simple as possible. Everything is here, from the elements of relativistic quantum mechanics to a mention of proton decay in grand unified models, with renormalization and even a brief discussion of the quark-parton model on the way.

Inevitably in a short book of wide scope, nothing is treated very fully. In places, plausibility arguments replace proper derivations. But experimentalists especially will find it helpful in providing an understanding of the essential ideas behind what the theorists are talking about, while neophyte theorists will find it a valuable introduction to the more advanced texts. The second half of the book will be the most useful, for the painless insight that it gives into non-Abelian physics. □

Peter Landshoff is Reader in Mathematical Physics at the University of Cambridge.